©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Numerische Klassifikation von Individuen und Merkmalsnormierung

(Numerical Classification of individuals and scaling of characters)

Von JOHANN HOHENEGGER

Mit 21 Abbildungen

(Vorgelegt in der Sitzung der mathem.-naturw. Klasse am 21. Jänner 1982 durch das w. M. H. ZAPFE)

Abstract

First it is shown that for an optimal classification analysis of a sample of individuals all discernable characters should be included. Equal weights of characters of different physical sizes and measures are achieved by scaling, so quantitative and qualitative characters can be used in numerical classification procedures based on the order of dissimilarities between individuals. Euclidean distance is the best metric measure of dissimilarity type and standardization of characters seems to be the most appropriate scaling-method as long as the frequency distribution of characters is normal, otherwise considerable distortion of the character space and the order of distances and similarity-relations result.

But if non-normal frequency distributions of characters are represented as sums of normal distributed components and each character is standardized with the distribution parameters of each of the components, the structure of characters posessing group-differences are enhanced in comparison to characters ill-suited for classification. The advantages of this new method of scaling in contrast to simple standardization is demonstrated with artificial and real data.

Key words: Numerical Classification, Morphometry, Characters, Scaling, Scatter-diagram, multivariate analysis.

Zusammenfassung

Werden Individuen miteinander verglichen, müssen in eine Klassifikation alle erfaßbaren Merkmale mit gleichem Gewicht eingehen. Diese gleiche Gewichtung wird bei Merkmalen unterschiedlicher physikalischer Größen bzw. mit differierenden Maßeinheiten durch eine Normierung erreicht. Sowohl quantitative als auch qualitative Merkmale können nach einer Normierung in eine numerische Klassifikationsanalyse eingehen, die auf der Ordnung von Ähnlichkeitsrelationen zwischen den Individuen basiert. Bei einer Darstellung der Merkmale als Achsen eines Euklidischen Raumes empfiehlt sich die Distanz als bestes metrisches Ähnlichkeitsmaß.

Von den Normierungsmethoden bringt die Standardisierung der Merkmalswerte die besten Ergebnisse, sie führt aber, wenn die Häufigkeitsverteilungen der Merkmale keiner Normalverteilung entsprechen, zu Verzerrungen des Merkmalsraumes. Dies wirkt sich auf die Distanzen und somit auf die Ähnlichkeitsrelationen der Individuen aus.

Werden die Häufigkeitsverteilungen der Merkmale in normalverteilte Komponenten zerlegt und erfolgt dann eine Standardisierung der Merkmalswerte mit den Verteilungsparametern der Komponenten, so wird eine Gewichtung der deutlicher strukturierten Merkmale gegenüber den unstrukturierten Merkmalen erreicht. Die Vorteile dieser Form der Normierung, im folgenden „komponentenspezifische Standardisierung" bezeichnet, gegenüber der einfachen Standardisierung, werden mit artifiziellen und realen Daten vorgeführt.

Klassifikation von Individuen und allgemeiner Klassifikationsbegriff

Die Gruppierung von Individuen auf Grund ihrer Ähnlichkeiten ist das elementare Problem der Klassifikation. Unter dem Begriff „Klassifikation" soll die Partitionierung einer Objektmenge in weitestgehend

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

homogene Teilmengen (= Gruppen, Cluster, Klassen) verstanden werden, wobei die Unterschiede zwischen den Gruppen deutlich sein sollen. Diese Definition unterscheidet sich von der biologischen (systematischen) Klassifikation, die sich auf das Einordnen von Organismen in Gruppen oder Reihen auf Grund ihrer Verwandtschaftsbeziehungen beschränkt (vgl. SIMPSON, 1961, 9).

Da die Klassifikation auf eine Objektmenge zurückgreift, bei der die homogenen Teilmengen erst ermittelt werden müssen, unterscheidet sie sich vom Prozeß der Identifikation, der sich auf das Ergebnis einer Klassifikation bezieht. Bei der Identifikation wird ein Individuum jener Teilmenge einer vorhandenen Klassifikation zugewiesen, zu der es die größte Ähnlichkeit besitzt. Beide Vorgänge – Klassifikation und Identifikation – wurden und werden in der Biologie häufig gemeinsam unter dem Sammelbegriff „Klassifikation" verstanden. Sowohl in der allgemeinen als auch in der biologischen Definition sind beide Prozesse jedoch deutlich verschieden.

Die meisten Klassifikationen werden in der Biologie und Paläontologie intuitiv durchgeführt. Da eine Gruppierung jedoch von der Geschicklichkeit und Sachkenntnis des Bearbeiters abhängt, führt dies zu einer subjektiven Interpretation der Wirklichkeit.

Als Beispiel möge die Untersuchung von BROLSMA, 1978, 57 ff., dienen, wo die gleiche Objektmenge, es handelt sich um Foraminiferen, vier verschiedenen Spezialisten zur Bestimmung vorgelegt wurde. Das Resultat erbrachte große Differenzen in der Klassifizierung. Während ein Bearbeiter nur 17 Arten, ein zweiter 32 und der dritte 36 Arten registrierte, konnte der vierte sogar 51 Arten feststellen. Dazu gab es kaum Übereinstimmungen in der Benennung dieser Gruppen. Es trat somit der Fall ein, daß sich die Spezialisten durch ihre subjektive Klassifizierung untereinander nicht mehr verständigen konnten.

Um die oben angeführten Probleme zu umgehen, wurde anfangs der fünfziger Jahre dieses Jahrhunderts der Versuch unternommen, den Klassifikationsprozeß zu „objektivieren" und die Güte der Klassifikation überprüfbar zu machen. Von der Biologie initiiert entwickelte sich ein selbständiger Wissenschaftszweig, der sich ausschließlich mit den Problemen der mathematischen Klassifikation befaßt und im Laufe seiner kurzen Geschichte zahlreiche Namen erhielt, die von Numerische Taxonomie (SOKAL & SNEATH, 1963) über Clusterung (HARTIGAN, 1975), Mathematische Taxonomie (JARDINE & SIBSON, 1971) und Numerische Klassifikation (CLIFFORD & STEPHENSON, 1975) bis zur Automatischen Klassifikation (BOCK, 1974) reichen. Durch diese mathematischen Gruppierungsmethoden entstanden jedoch neuerdings zahlreiche Probleme. Auf einige von ihnen soll in der folgenden Arbeit eingegangen werden. Doch vorher müssen der Begriff Klassifikation und die damit verbundenen Methoden näher erläutert werden.

Jedes Individuum – oder allgemeiner Objekt – läßt sich mit einer bestimmten Zahl von Merkmalen und deren Ausprägungen charakterisieren. Anhand der Koinzidenzen von Merkmalswerten lassen sich Ähnlichkeiten zwischen den Individuen definieren. Die Merkmale

können sowohl morphologischer als auch physiologischer, ökologischer, ethologischer und geographischer Natur sein. Die einmal definierten Ähnlichkeiten dienen als Kriterium der Klassifikation. Treten innerhalb einer Objektmenge Gruppen ähnlicher Organismen auf, und sind diese Gruppen untereinander verschieden, so ist eine innere Struktur der Objektmenge gegeben, und die Klassifikation läßt sich durchführen. Man kann an dieser Stelle den Begriff der natürlichen Klassifikation einführen und ihm den Begriff einer künstlichen Klassifikation gegenüberstellen:

„Eine Klassifikation ist natürlich, wenn hinsichtlich der Merkmalsausprägungen innerhalb einer Gruppe Kontinuität und zwischen den Gruppen Diskontinuitäten bestehen. Sie ist künstlich, wenn hinsichtlich der Merkmalsausprägungen innerhalb einer Gruppe Diskontinuitäten oder zwischen den Gruppen Homogenitäten aufscheinen (vgl. dazu Simpson, 1961, 114 ff.; Blackwelder, 1967, 186)."

In der systematischen Biologie werden durch den Prozeß der Klassifikation phänotypisch ähnliche Individuen zusammengefaßt, eine solche Gruppe wird von Mayr, 1975, 16, als „Phänon" bezeichnet. Phänotypisch ähnliche Klassen müssen aber nicht unbedingt auf verwandtschaftlichen Beziehungen beruhen. Durch Konvergenzen und Parallelentwicklungen können analoge Strukturen entwickelt werden, die Ähnlichkeiten zwischen phylogenetisch entfernten Gruppen bewirken können. Somit ergeben sich bei den höheren systematischen Kategorien die Differenzen zwischen dem biologischen und dem allgemeinen Klassifikationsbegriff. Trotzdem bringt eine rein auf dem phänetischen Prinzip basierende Klassifikation auch bei den höheren systematischen Kategorien eine gute Annäherung an die phylogenetische Klassifikation (nicht im Sinne von Hennig, 1966) (vgl. dazu Jardine & Sibson, 1971, 140 ff.).

Im Artniveau und darunter, wo innerhalb einer Lokalpopulation (= Mendel-Population) analoge Merkmale bei den Individuen auf Grund der Vererbung nicht auftreten können, gibt es keine Differenzen zwischen der biologischen und allgemeinen Klassifikation. Die Ähnlichkeit zwischen den Individuen ist durch die Vererbung bewirkt, da die Anlagen zu den Merkmalsausprägungen an die Nachkommen weitergegeben werden und in den Fällen der geschlechtlichen Fortpflanzung immer neue Kombinationen dieser Erbanlagen möglich sind. Die Homogenität in den Merkmalen einer Organismengruppe wird durch nahe verwandte, oft abstammungsgleiche Gene hervorgerufen.

Eine mehr oder minder festgelegte Zahl von Erbfaktoren bedingt in ihren verschiedenen Kombinationen die genetische Variabilität einer biologischen Gruppe. Die Variabilität eines Merkmals wird aber auch noch durch die Umwelt beeinflußt. Umweltbedingte Variationen, sogenannte Modifikationen, besitzen jedoch ihrerseits wiederum einen genetischen Rückhalt, da die Erbanlagen die Grenzen der Reaktionsnorm bestimmen.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Da die phänetische Ähnlichkeit von Individuen ausschließlich durch die abstammungsverwandten Erbanlagen und keine anderen Ursachen hervorgerufen wird, kann man daraus schließen, daß Ähnlichkeiten von Individuen die genetischen Beziehungen (= Verwandtschaft) widerspiegeln. Im Bereich des Artniveaus und darunter bestehen somit keine Widersprüche zwischen dem biologischen und dem allgemeinen Klassifikationsbegriff.

Bisher wurde von den Vertretern der Numerischen Taxonomie größeres Augenmerk auf die höheren systematischen Kategorien gerichtet und die Klassifizierung von Individuen vernachlässigt. Dies geschah teils aus theoretischen Überlegungen (vgl. JARDINE & SIBSON, 1971, deren Methoden an Gruppen von Individuen durchgeführt werden), teils aus praktischen Gründen (vgl. SNEATH & SOKAL, 1973), weil damals die Kapazität der Rechenmaschinen eine große Rolle spielte.

Das Problem der Klassifikation von Individuen ist in der modernen biologischen Systematik tatsächlich nicht wesentlich, da die Systematiker auf biologische Populationen zurückgreifen können und hier die Möglichkeit besitzen, neben den morphologischen auch alle anderen taxonomischen Merkmale zu erfassen.

In der Paläontologie stellt sich jedoch folgendes Problem: Hier liegen in den meisten Fällen Mischungen von Mitgliedern mehrerer biologischer Populationen vor, die nur in den seltensten Fällen gleichzeitig gelebt hatten und die außerdem nur selten in ihrem ursprünglichen Lebensraum erhalten sind. Es bleiben für eine Klassifikation meist nur morphologische Merkmale übrig. Eine ,,Fossilpopulation" ist also keineswegs mit einer biologischen Population vergleichbar, bei der sich ähnliche Individuen sofort als biologische Einheit ansprechen lassen. In einer Fossilpopulation können, was rezent selten und nur unter bestimmten Umständen vorkommt (vgl. SOLBRIG & SOLBRIG, 1979, 265 ff.; WHITE, 1978, 227 ff.), mehrere nahe verwandte Formen, die morphologisch nur gering divergieren, vorkommen. Diese Divergenzen müssen durch Klassifikationsprozesse erfaßt werden, damit die Unterschiede zwischen den Gruppen verdeutlicht werden können. Als eine Forderung an eine solche Klassifikation muß jedoch die Überprüfbarkeit gestellt werden, die im Falle einer intuitiven Gruppierung nicht gewährleistet ist. Deshalb müßte den automatischen Klassifikationsverfahren der Vorrang eingeräumt werden.

Numerische Klassifikation

Sollen Individuen klassifiziert werden, muß man zuerst die Merkmale festlegen, anhand derer Ähnlichkeitsbestimmungen zwischen den Individuen durchzuführen sind. Ein Individuum hat für jedes Merkmal einer Merkmalsmenge einen bestimmten Wert aufzuweisen, der als Merkmalsausprägung bezeichnet wird (vgl. FERSCHL, 1978, 17).

Mit diesen Merkmalsausprägungen der Individuen einer Objektmenge lassen sich Untersuchungen über die Variabilität der Merkmale

anstellen. Anhand dieser Untersuchungen können Diskontinuitäten zwischen Gruppen, die ihrerseits homogen sind, festgestellt werden. Dies gilt für alle Arten von Merkmalen.

Bei sogenannten „qualitativen" oder klassifikatorischen Merkmalen, bei denen bereits eine Klassifikation vorgegeben ist (z. B. männlich – weiblich, vorhanden – nicht vorhanden, rot – blau – grün – gelb usw.), ist die Variabilität innerhalb einer Population dann gegeben, wenn die Merkmalsalternativen durch genetischen Polymorphismus hervorgerufen werden. Ist dies nicht der Fall, hätten diese Merkmale ein großes klassifikatorisches Gewicht, da die vorgegebene Klasseneinteilung des Merkmales bereits die „natürliche" (biologische) Klassifikation widerspiegelt. Vor einem Klassifikationsprozeß lassen sich aber die Ursachen dieser Variabilität (in unserem Fall inter- oder intraspezifische Kausalitäten) nicht determinieren, sie können erst durch das Ergebnis einer Klassifikation interpretiert werden.

Ähnliche Überlegungen gelten für die topologischen Merkmale. Hier stehen die einzelnen Merkmalsalternativen in einer Ordnungsrelation zueinander, die Intervalle zwischen den Merkmalsklassen bleiben unberücksichtigt.

Metrische (= „quantitative") Merkmale besitzen den höchsten Informationsgehalt aller Merkmalsarten, da auch die Intervalle zwischen den Merkmalsausprägungen klassifikatorisch ins Gewicht fallen.

Aus diesen Gründen sind die besten Variationsuntersuchungen mit metrischen Merkmalen durchzuführen, und in der Folge soll hauptsächlich auf sie zurückgegriffen werden. Will man die Variabilität der Merkmale im Hinblick auf Diskontinuitäten untersuchen, so nimmt man im einfachsten Fall ein Merkmal und untersucht die Häufigkeitsverteilung dieses Merkmals in der Stichprobe der Individuen. Ergibt sich das Bild einer eingipfeligen Verteilungsform (Abb. 1, Histogramm bei der Abszisse des Streuungsdiagramms), läßt sich der Schluß ziehen, daß die gesamte Objektmenge hinsichtlich dieses Merkmals homogen ist. Eine Überprüfung der Homogenität läßt sich mit verschiedenen statistischen Verfahren, wie beispielsweise Chi-Quadrat-, Kolmogoroff-Smirnov- oder G-Test (vgl. Sokal & Rohlf, 1969), durchführen. Im Beispiel der Abb. 1 liegen keine Gründe zur Unterteilung dieser Objektmenge vor.

Würde sich jedoch das Bild einer zwei- oder mehrgipfeligen Verteilungsform ergeben, kann die Annahme getroffen werden, daß mehrere homogene Verteilungen eine Mischverteilung bewirken und die Gipfelpunkte (Modalwerte) die Lageparameter der einzelnen Teilgruppen andeuten. Eine Aufgliederung solcher Mischverteilungen in Teilgruppen läßt sich mit verschiedenen Methoden durchführen (vgl. Medgyessy, 1977).

Ergeben sich jedoch wie in unserem Beispiel der Abb. 1 keine Hinweise auf Inhomogenitäten und Teilmengen, ist man versucht, dieses Merkmal als für die Klassifikation unwichtig zu werten. Man kann in der Folge ein weiteres Merkmal heranziehen und dieses auf Inhomogenitäten

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

in der untersuchten Objektmenge überprüfen (vgl. Abb. 1, Histogramm bei der Ordinate des Streuungsdiagrammes). Wenn nun auch beim zweiten und bei den folgenden Merkmalen keine Inhomogenitäten auftreten, ist man noch keineswegs berechtigt, die Objektmenge als homogen, d. h. nicht klassifizierbar, zu bezeichnen. Es blieben bei dieser Vorgangsweise bisher nämlich alle Merkmalskombinationen unberück-

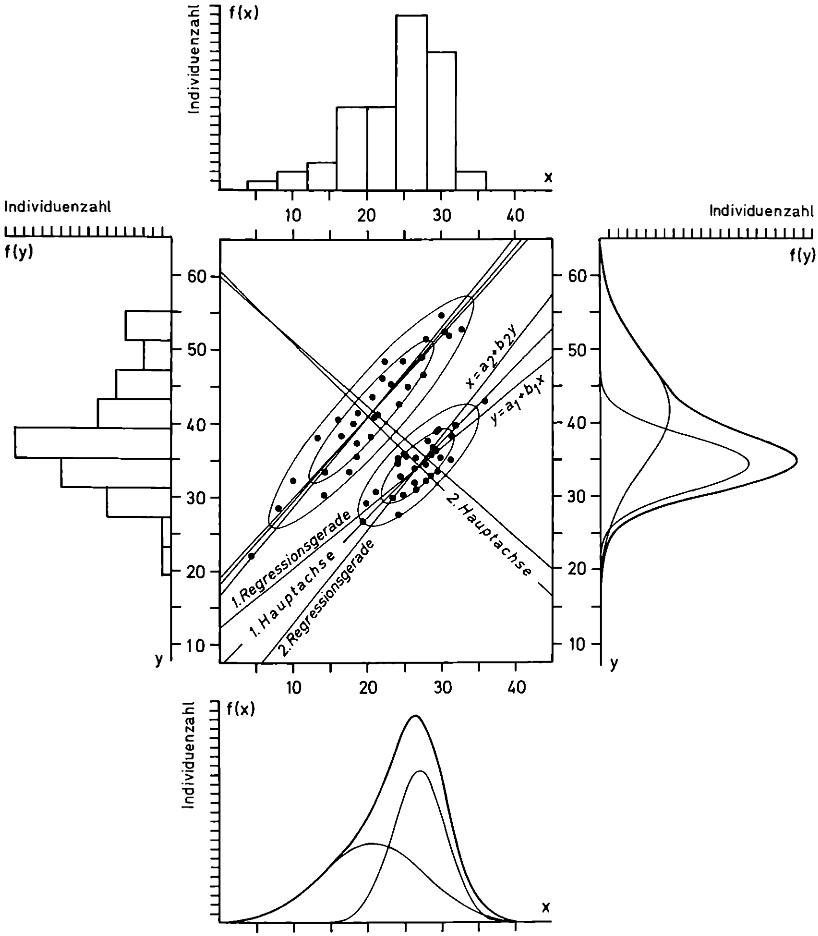


Abb. 1: Konfiguration von 60 Individuen im zweidimensionalen Euklidischen Merkmalsraum (= Streuungsdiagramm). Die beiden Klassen (= Gruppen) zu je 30 Individuen sind in den Merkmalen x und y jeweils bivariat normalverteilt. Ihre Streuungsbereiche sind durch Ellipsen gekennzeichnet. Innerhalb dieser Ellipsen kommen 60 bzw. 90 % der Individuen einer Klasse zu liegen. Weiters sind die Regeressionsgeraden und Hauptachsen der Verteilungen eingetragen. Die empirischen und theoretischen Häufigkeitsfunktionen der einzelnen Merkmale – einerseits (empirisch) Histogramme (= Säulendiagramm), andererseits (theoretisch) Summenkurven von Normalverteilungen – sind als Randverteilungen des Streuungsdiagrammes dargestellt.

sichtigt. Das Streuungsdiagramm der Abb. 1 soll verdeutlichen, wie eine solche Erfassung von Merkmalskombinationen möglich ist. Hier werden die zwei Merkmale, bei denen in den Häufigkeitsverteilungen keine Inhomogenitäten auftreten, zueinander in Beziehung gesetzt, indem man sie als Koordinatenachsen eines zweidimensionalen Raumes betrachtet (man spricht dann vom sogenannten Merkmalsraum). Mathematisch läßt sich dieser Merkmalsraum durch Einbeziehung weiterer Merkmale auf zahlreiche Dimensionen erweitern. Bei einer Zahl von n-Merkmalen spricht man von einem n-dimensionalen Merkmalsraum.

In einem euklidischen Merkmalsraum stehen für gewöhnlich die Achsen senkrecht aufeinander (Kartesische Koordinaten), was bedeutet, daß sie frei gewählt sind und keine Beziehung zueinander haben.

Dies ist jedoch eine Forderung, die besonders bei biologischen Merkmalen nicht zu stellen ist. Hier können in den Merkmalen einer Population zahlreiche Korrelationen auftreten (z. B. Körpergröße und Gewicht). Diese Korrelationen lassen sich eliminieren, wenn man die Hauptachsen (vgl. Blackith & Reyment, 1971, 146 ff.) der multidimensionalen Punktwolke bestimmt und die Hauptachsen als Koordinatenachsen determiniert. Diese Achsen liegen in Richtung der größten Streuung bzw. Reststreuung und stehen normal aufeinander.

Im zweidimensionalen Merkmalsraum sind die Unterschiede zwischen dem ursprünglichen Merkmalsraum und dem Raum, der von den Hauptachsen aufgespannt wird, nur gering. Bei Verwendung vieler Merkmale, also im multivariaten Merkmalsraum, treten jedoch oft sehr starke Korrelationen auf. Hier gewinnt die Hauptachsentransformation große Bedeutung.

Im Streuungsdiagramm der Abb. 1, bei einer Kombination der beiden Merkmale, zeigt sich sehr wohl eine Diskontinuität innerhalb der bivariaten Verteilung. Man kann zwei Gruppen (Klassen) unterscheiden, die ihrerseits homogen sind. Die Randverteilungen, wie die eindimensionalen Häufigkeitsverteilungen der Merkmale bei einer multivariaten Darstellung genannt werden (vgl. Kreyszig, 1968, 144 ff.), zeigen in unserem Beispiel zwar statistisch überprüfbare Homogenitäten, in der Merkmalskombination ergeben sich jedoch Diskontinuitäten, die zu einer Klassifizierung berechtigen. Dies bedeutet, daß trotz Bestätigung durch einen statistischen Homogenitätstest die Realität ganz anders geartet sein kann und deshalb immer eine äußerst vorsichtige Analyse gegebener Daten geboten ist.

Auf den allgemeinen Fall bezogen, muß man bei einer Klassifikation von Individuen alle erfaßbaren Merkmale, sowohl quantitative als auch qualitative, heranziehen und sie auch als multidimensionale Merkmalskombinationen betrachten, um eventuelle Diskontinuitäten entdecken zu können.

Diese Erkenntnis der Bedeutung von Merkmalskombinationen für die Taxonomie der höheren systematischen Kategorien kam nicht von seiten der Numerischen Taxonomie, sondern wurde schon früher erkannt

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

(vgl. SIMPSON, 1961, 41 ff.; MAYR, 1975, 81). Nimmt man bei einer biologischen systematischen Klassifikation jedes Merkmal für sich, trifft man die Annahme, daß das einzelne Merkmal das Taxon absolut zu kennzeichnen vermag. Für ein solches Klassifikationskonzept wurde der Begriff monothetisch eingeführt.

Bereits ADANSON (nach MAYR, 1975) hat 1763 erkannt, daß die meisten höheren Taxa durch eine Merkmalskombination gekennzeichnet sind, wobei jeder Vertreter des untersuchten Taxons die Mehrheit der Merkmale aufweist. Man spricht in diesem Fall von einer polythetischen Klassifikation.

Das Beispiel in Abb. 1 und die vorangegangenen Ausführungen zeigen, daß man auch bei der Gruppierung von Individuen von einer monothetischen und polythetischen Klassifikation sprechen kann.

Jeder Systematiker, der Individuen klassifizieren möchte und dazu nicht auf die Gesamtzahl der erfaßbaren Merkmale, sondern nur auf wenige zurückgreift, mißt diesen a priori größeres Gewicht bei. Er begibt sich damit in die Gefahr, die natürlichen Klassen nicht zu erfassen. Diese Vorgangsweise, auch wenn sie sich durch die Anwendung statistischer Methoden den Anschein von Exaktheit gibt, unterscheidet sich kaum von der subjektiven Klassifizierung, umso mehr, da der intuitive Systematiker bei der Betrachtung der Objekte eine Fülle von Merkmalen unbewußt wahrnimmt und sie bei seiner Klassifikation verwendet, obwohl sie dann bei der Charakterisierung der Gruppen nicht aufscheinen.

Nachdem gezeigt wurde, daß in eine Klassifikation von Individuen alle verfügbaren Merkmale und deren Kombinationen eingehen müssen, stellt sich das Problem einer numerischen Erfassung von Merkmalsausprägungen, ohne die eine automatische Klassifikation undurchführbar wäre. Dieses allgemeine Problem der Meßtheorie wird als Skalierung bezeichnet. Eine Skala ist eine relationstreue Abbildung eines Gegenstandsbereiches in ein System von reellen Zahlen (vgl. FERSCHL, 1978, 21). Bei einer Skalierung wird der Versuch unternommen, empirisch erfaßte Beziehungen in eine numerische Relation umzusetzen.

Messungen der oben angeführten Merkmalstypen (klassifikatorische, topologische und metrische Merkmale) lassen sich in spezielle Skalen transformieren, die sich in Art und Zahl der zulässigen mathematischen Operationen unterscheiden (Nominal-, Ordinal- und Intervall- bzw. Rationalskala; vgl. AHRENS, 1974, 63 f.). Während für metrische Merkmale die Überführung in Skalen keine Schwierigkeiten bereitet, trifft dies bei den beiden anderen Typen von Merkmalen nicht zu. Hier muß man bei einer Skalierung zwischen binären Merkmalen (z. B. vorhanden – nicht vorhanden) und mehrstufigen Alternativmerkmalen unterscheiden (vgl. BOCK, 1974, 66 ff.). Diese können ungeordnete (z. B. Farben) oder geordnete Alternativen (z. B. Güteklassen) sein.

Binäre Merkmale werden oft durch die Zahlen 0 und 1 dargestellt, es handelt sich dabei aber um keine Skala im eigentlichen Sinn, da zwischen

den Werten keinerlei Relationen bestehen (man könnte auch andere Symbole verwenden, die keine Zahlen sind, wie beispielsweise + und –).

Für geordnete Alternativmerkmale, die in Relation zueinander stehen, ist – wenn es nur um die Erhaltung der Ordnungsrelation geht – jede Zahlenreihe gültig, die diese Ordnung der Merkmalsausprägungen widerspiegelt. Meist werden sie mit der Menge der natürlichen Zahlen symbolisiert, es ist aber jede Skala zulässig, die die Ordnungsrelation enthält (z. B. 1, 4, 9, 16 usw.).

Ein mehrstufiges Alternativmerkmal mit ungeordneten oder geordneten Alternativen läßt sich auch durch mehrere binäre Merkmale darstellen; deren Anzahl hängt von der Zahl der Stufen des mehrstufigen Merkmals ab (vgl. SNEATH & SOKAL, 1973, 147 ff.). Die Methode, Alternativmerkmale in eine Reihe binärer Skalen überzuführen, ist elementarer als die Skalierung und wird als Kodierung bezeichnet (SNEATH & SOKAL, 1973, 147).

Bei metrischen Merkmalen ergeben sich die geringsten Schwierigkeiten in deren Skalierung. Hier sind die Skalen entweder durch Messungen festgelegt (direkte Messungen) oder die Merkmalsausprägungen stellen absolute Werte dar (indirekte Messungen, wie z. B. Indizes, Quotienten).

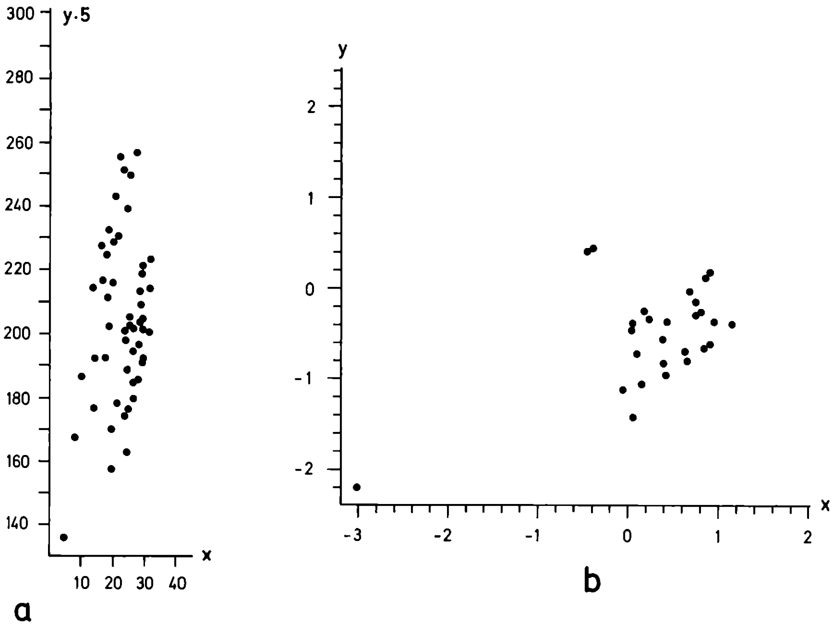


Abb. 2: Verzerrungen des Merkmalsraumes der Abb. 1.
a) Verzerrung durch Änderung der Maßeinheiten des Merkmals y auf 1/5 der ursprünglichen Skala. Die deutliche Trennung der beiden Klassen, wie sie in Abb. 1 dargestellt ist, geht durch diese Transformation verloren.
b) Verzerrung durch Standardisierung. Die Randverteilungen der beiden Merkmale streuen symmetrisch um den Mittelwert 0. Durch die annähernd gleiche Dimensionierung der Wertebereiche beider Merkmale wird die Struktur der Objektmenge, die Trennung der beiden Klassen, deutlich sichtbar.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Eine Objektmenge erfaßt mit den metrischen Werten ihrer unterschiedlichen Merkmalsausprägungen nur einen bestimmten Abschnitt der Skala, der als Wertebereich bezeichnet wird. Dieser Wertebereich hängt bei direkten Messungen sehr stark von den Maßeinheiten ab, in denen das Merkmal gemessen wird. Diese Unterschiedlichkeit in den mehr oder minder willkürlich gewählten Maßeinheiten kann zu großen Verzerrungen im Merkmalsraum führen. Die Abbildung 2a möge dies verdeutlichen. Hier wurde das zweite Merkmal der Abb. 1 in einer Maßeinheit dargestellt, die 1/5 der Maßeinheit des ersten Merkmals ist. Die Darstellung im Euklidischen Raum zeigt deutlich die starke Verzerrung der Gruppenstrukturen. In diesem Fall sind die beiden deutlich getrennten Gruppen der Abb. 1 nicht mehr auseinanderzuhalten, obwohl hier nur eine einfache und für Daten auf dem Niveau einer Rationalskala zulässige Transformation der Form $y' \rightarrow ay$ bei einer Variablen durchgeführt wurde.

Das Problem der Ungleichwertigkeit von Skalen in den Merkmalsbereichen läßt sich dadurch umgehen, daß man jedes Merkmal normiert. Die Wertebereiche der einzelnen Variablen werden auf annähernd gleiche Länge gebracht. Man spricht dann von normierten Daten, die dimensionslos sind. Durch eine Normierung kommt den Merkmalswerten größenmäßig die gleiche Bedeutung zu.

Es haben sich zwei Methoden der Normierung durchgesetzt. Zuerst soll die „Normierung auf den Merkmalsbereich" erläutert werden (Intervallgrenzen 0 und 1). Diese Normierung (vgl. SNEATH & SOKAL, 1973, 153; BOCK, 1974, 37) hängt von den Extremwerten einer Verteilung ab, da von jedem Merkmalswert der kleinste Merkmalswert subtrahiert und dann durch die Variationsbreite dividiert wird:

$$x^*_{ik} = \frac{x_{ik} - \text{Min}\,[x_{ik}]}{\text{Max}\,[x_{ik}] - \text{Min}\,[x_{ik}]} \qquad (1)$$

$i = 1, \ldots n$ (n = Zahl der Individuen);
$k = 1, \ldots m$ (m = Zahl der Merkmale).

Der Nachteil bei dieser Art der Normierung besteht darin, daß sie nur von zwei Merkmalswerten, den beiden Extremwerten einer Verteilung, abhängt, die in den Stichproben stärksten Schwankungen unterliegen können, wie es das Beispiel Tab. 1 verdeutlichen soll. In diesem Beispiel wurden aus einer gegebenen Grundgesamtheit fünf Stichproben gezogen, von denen drei denselben Umfang hatten. Eine Stichprobe ist kleiner, die andere größer als die übrigen Stichproben. Betrachtet man die Variationsbreiten, so sind die Unterschiede sehr groß. In der Abb. 3 wurde der Versuch unternommen, diese Abweichungen der Stichproben von der Grundgesamtheit graphisch darzustellen. Eine Grundgesamtheit von 50 Objekten wurde auf den Merkmalsbereich (Intervallgrenzen 0 und 1) normiert und mit den normierten Stichproben

verglichen. Die Positionen der Objekte auf der Zahlengeraden stimmen in der Grundgesamtheit und Stichprobe nur dann überein, wenn die Stichproben die extremen Individuen der Grundgesamtheit enthalten (vgl. Abb. 3, 5. Stichprobe). In allen anderen Fällen sind keine Übereinstimmungen gegeben, die Abweichungen können oft extreme Ausmaße annehmen (vgl. Abb. 3, 4. Stichprobe). Außerdem ergeben sich beim Vorhandensein von sogenannten „Ausreißern" in den Stichproben große Probleme, da dann die Daten der gesamten Stichprobe mit einem falschen Wert normiert werden und somit weitere Auswertungen völlig sinnlos machen können.

Tabelle 1: Vergleiche der Verteilungsparameter von Stichproben mit denen der Grundgesamtheit (siehe Abb. 3).

	Individuenzahl	Arithmetisches Mittel	Standardabweichung	Minimum	Variationsbreite
Grundgesamtheit	50	30,97	9,67	8,6	44,3
1. Stichprobe	10	30,11	6,83	13,5	24,7
2. Stichprobe	10	34,24	11,38	16	32,8
3. Stichprobe	10	32,85	9,27	18,9	28,9
4. Stichprobe	5	27,32	10,82	16,2	20,1
5. Stichprobe	15	29,33	10,73	8,6	44,3
Streuungen (Standardabweichungen) der Stichprobenparameter		2,77	1,83	3,88	9,21

Da diese Form der Normierung zu stark von wenigen Merkmalswerten abhängt, ist es besser, eine Form der Normierung zu verwenden, die alle Individuen einer Stichprobe im gleichen Maße berücksichtigt. Als statistisch geeignete Normierung bietet sich die Standardisierung an. Bei einer Standardisierung wird von jedem Merkmalswert das arithmetische Mittel der Stichprobe subtrahiert und durch die Standardabweichung dividiert:

$$x^*_{ik} = \frac{x_{ik} - \bar{x}_k}{s_k} \qquad (2)$$

$\bar{x}_k$ = arithmetisches Mittel des k-ten Merkmals der Objektmenge;
s_k = Standardabweichung des k-ten Merkmals der Objektmenge

$$= \left[\frac{1}{n}\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2\right]^{1/2}$$

Durch diese Form der Normierung erhält jedes Merkmal eine Verteilung mit dem arithmetischen Mittel 0 und der Standardabweichung 1. Auch hier läßt sich ein Vergleich der Positionen von Individuen zwischen Grundgesamtheit und Stichprobe anstellen. Schon aus Tabelle 1

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

ist zu ersehen, daß sowohl die Lageparameter (arithmetische Mittel) als auch die Streuungsmaße (Standardabweichungen) bei den Stichproben wesentlich geringer variieren als die äquivalenten Normierungsparameter der Merkmalsbereichsnormierung (Minimum und Variationsbreite). Die Abb. 3 zeigt den Vergleich zwischen standardisierter Grundgesamtheit und den standardisierten Stichproben. Hier erweist es sich, daß das arithmetische Mittel und die Standardabweichung einer Stichprobe tatsächlich gute Schätzgrößen für die äquivalenten Parameter der Grundgesamtheit sind. Die Differenzen der Positionen auf der Zahlengeraden zwischen Stichprobe und Grundgesamtheit sind bei weitem nicht so groß wie jene, die durch eine Normierung auf den Merkmalsbereich hervorgerufen werden.

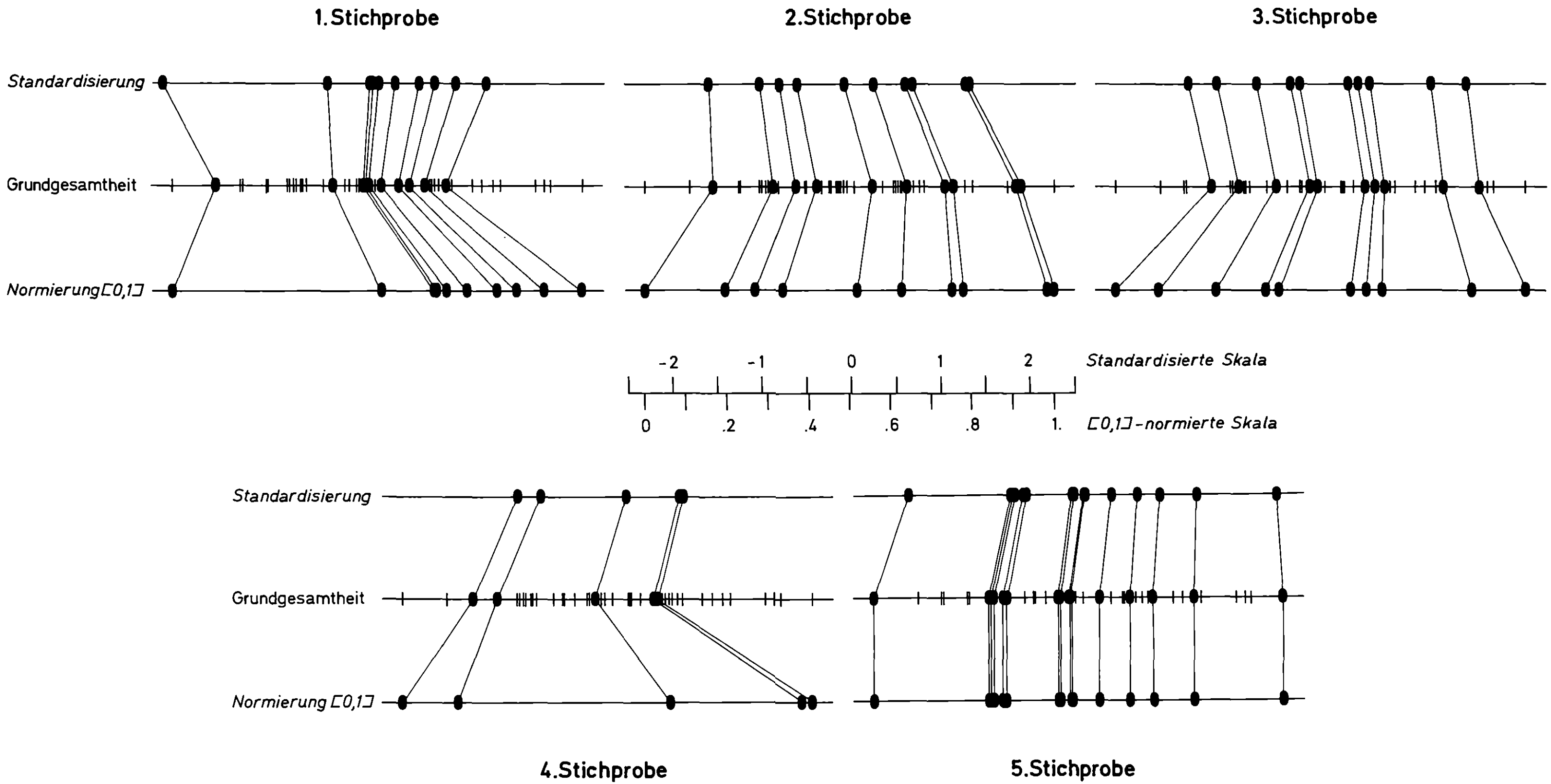
Durch die Standardisierung der Merkmalswerte werden alle Merkmale auf eine einheitliche Größe, in diesem Fall auf Verteilungen mit gleichen arithmetischen Mitteln und Standardabweichungen genormt. Die Wertebereiche hängen nicht mehr von den Maßeinheiten ab. Das Ergebnis der Standardisierung von Merkmalsausprägungen der Objekte in Abb. 1 wird in Abb. 2 b dargelegt. Durch die gleiche Gewichtung der Merkmalsvarianzen gewinnen die Wertebereiche ungefähr gleiche Bedeutung, und Inhomogenitäten in den Verteilungen können deutlich zutage treten. Nimmt man den ursprünglichen Merkmalsraum der Abb. 1, wo die zweite Variable eine größere Variationsbreite aufweist, so treten durch die Standardisierung gleichfalls Verzerrungen des Merkmalsraumes auf. Auf diese durch die Standardisierung hervorgerufenen Probleme soll im anschließenden Kapitel ausführlich eingegangen werden. Hier ist noch zu bemerken, daß durch solche Verzerrungen besonders die Ähnlichkeitsrelationen zwischen den Individuen betroffen werden.

Wie sind aber die Ähnlichkeiten im euklidischen Merkmalsraum zu definieren? Nimmt man einen ursprünglichen Merkmalsraum, wie beispielsweise den zweidimensionalen der Abb. 1, so ergeben sich durch die gleichmäßige Abdeckung von Streuungsbereichen räumliche Homogenitäten innerhalb der beiden Gruppen. Die Inhomogenität zwischen den Klassen ist durch einen Zwischenraum gekennzeichnet, in dem kein Individuum mit seinen Merkmalsausprägungen zu liegen kommt. In der Darstellung der Individuen als Objekte im mehrdimensionalen Euklidi-

---

Abb. 3: Vergleich von Standardisierung und Normierung auf den Merkmalsbereich (Intervallgrenzen 0 und 1).
Die Positionen der Individuen einer standardisierten bzw. auf den Merkmalsbereich normierten Grundgesamtheit von 50 Individuen sind auf einer Zahlengeraden dargestellt (mittlere Gerade bei den einzelnen Stichproben; Maßeinheiten in der Mitte der Abbildung). Durch Standardisierung und Normierung auf den Merkmalsbereich der Individuen mit den Verteilungsparametern der fünf Stichproben (vgl. Tab. 1) ergeben sich Abweichungen von den tatsächlichen Positionen auf der Zahlengeraden, wobei im Gegensatz zu den auf den Merkmalsbereich normierten Daten die standardisierten Werte die geringeren Differenzen zu den tatsächlichen Werten der normierten Grundgesamtheit zeigen.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at



©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

schen Raum, die bei den meisten multivariaten Klassifikationsmethoden Verwendung findet, bestimmen die Distanzen zwischen den Objekten (z. B. Euklidische Distanz) den Grad der Ähnlichkeit (vgl. Abb. 6):

$$D_{ij} = \left[\sum_{k=1}^{m} (x_{ik} - x_{jk})^2 \right]^{1/2} \qquad (3)$$

D_{ij} = Euklidische Distanz zwischen den Individuen i und j.

Die Position der Individuen im Merkmalsraum lassen sich, ohne auf die Koordinatenwerte der einzelnen Individuen zurückgreifen zu müssen, durch die Distanzen einwandfrei bestimmen. Nimmt man Winkelmessungen zwischen den Individuen als Ähnlichkeitsmaße (z. B. Cosinus-Koeffizient, vgl. JÖRESKOG, KLOVAN & REYMENT, 1976, 89; Produktmomentkorrelationskoeffizient), kann man eine unendliche Zahl ähnlicher Merkmalsräume definieren, in denen nur die relative Lage der Individuen zueinander erhalten ist.

In der Regel werden bei der Mehrzahl der multivariaten Klassifikationsmethoden, wenn sie nicht direkt auf die Merkmalswerte zurückgreifen (z. B. Gradientenverfahren von SCHNELL, 1964), vorerst die Ähnlichkeiten zwischen den Individuen ermittelt und dann mit diesen ein Gruppierungsverfahren durchgeführt.

Auf die einzelnen automatischen Klassifikationsmethoden und deren Vor- bzw. Nachteile soll hier nicht näher eingegangen werden. Der interessierte Leser kann sich in zahlreichen Lehrbüchern darüber informieren (z. B. BOCK, 1973; HARTIGAN, 1975; JARDINE & SIBSON, 1971; SNEATH & SOKAL, 1973; VOGEL, 1975). Außerdem stehen für die rechnerischen Analysen fertige Programmpakete zur Verfügung (z. B. ROHLF, KISHPAUGH & KIRK, 1977; WISHART, 1978).

Wichtig ist hier zu erwähnen, daß von den biologisch orientierten Taxonomen, die automatische Verfahren anwenden, bei der Klassifikation von Individuen folgende Vorgangsweise empfohlen wird:

Innerhalb einer homogenen Klasse von Organismen ist hinsichtlich der Merkmalsausprägungen von Individuen bei diesen keine hierarchische Struktur gegeben, infolgedessen dürfen auch keine hierarchischen Klassifikationsmethoden angewendet werden (vgl. JARDINE & SIBSON, 1971, 163). Die mehrdimensionalen Ähnlichkeitsbeziehungen der Individuen werden bei Anwendung eines hierarchischen Verfahrens nicht im vollen Ausmaß genützt. Es wird daher die Anwendung einer multidimensionalen Skalierung empfohlen.

Bei der multidimensionalen Skalierung werden die Individuen als Punkte in einem Euklidischen Raum dargestellt, der neben einer geringen Dimensionalität (zwei bis maximal vier Dimensionen) die sogenannte Monotoniebedingung erfüllen muß. Diese besagt folgendes: Wenn im

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

ursprünglichen Merkmalsraum die Ähnlichkeiten zwischen den Individuen i und j größer sind als zwischen k und l, so soll diese Relation auch im transformierten Merkmalsraum gelten:

$D_{ij} > D_{kl} \Rightarrow d_{ij} > d_{kl}$.

D_{ij} = die Ähnlichkeit zwischen den Individuen i und j im hochdimensionalen Raum;

d_{ij} = die Ähnlichkeit im reduzierten Raum.

In der multidimensionalen Skalierung unterscheidet man metrische und nichtmetrische Methoden. Während in einer metrischen multidimensionalen Skalierung die Ähnlichkeiten in einer metrischen Skala dargestellt werden (vgl. GOWER, 1966; TORGERSON, 1958), genügen für die nichtmetrischen Methoden die Ordnungsrelationen der Ähnlichkeiten (vgl. SHEPARD, 1962a, b; KÜHN, 1976). In der biologischen Systematik findet man für die metrische Form der multidimensionalen Skalierung oft die Bezeichnung Hauptkoordinatenanalyse (z. B. BLAKKITH & REYMENT, 1971; SNEATH & SOKAL, 1973).

Die Individuen werden nach einem multidimensionalen Skalierungsverfahren als Punkte in einem Euklidischen Raum von geringer Dimensionalität dargestellt. Diskontinuitäten in der Punkteverteilung müssen subjektiv erfaßt werden. Um auch diesen Vorgang zu „objektivieren", werden verschiedene Wege vorgeschlagen. Beispielsweise könnte auf eine multidimensionale Skalierung eine Analyse folgen, die dem Klassifikationszwang unterliegt, d. h., die auf alle Fälle Klassen schafft (Clusteranalysen). Einerseits könnte man eine solche Analyse mit den Koordinatenwerten der Skalierungsachsen durchführen, andererseits könnte auch auf die ursprünglichen Merkmalswerte zurückgegriffen werden. Es müßten aber Verfahren eingesetzt werden, die überlappende Klassen erlauben; ein Individuum sollte also mehr als einer Klasse angehören können (vgl. JARDINE & SIBSON, 1971, 59ff.). Dadurch unterscheiden sich diese Verfahren von den hierarchischen Methoden, bei denen ein Individuum stets nur einer Klasse zugeordnet wird.

Der Nachteil der Analysen, die überlappende Klassen ermöglichen, besteht darin, daß ihre praktische Durchführung kompliziert ist. Daher kann meist nur eine geringe Zahl von Individuen (< 70) in eine solche Analyse eingehen. In der Zukunft könnte die immer größere Kapazität der modernen elektronischen Rechenanlagen es ermöglichen, daß auch eine große Individuenzahl, wie sie bei Populationen kleiner Individuen oder Mikroorganismen auftritt, in einem überlappenden Klassifikationsverfahren behandelt werden kann.

Normierung

Durch eine Normierung bekommen die Merkmale gleiches Gewicht, die relativen Unterschiede in der Variabilität gehen dadurch verloren. Vergleicht man den originalen (Abb. 1) mit dem standardisierten Merkmalsraum (Abb. 2b), erkennt man eine durch die Standardisierung

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

hervorgerufene geringe Verzerrung des Euklidischen Raumes. Angenommen, die beiden Merkmale der Abb. 1 wären in gleichen Skaleneinheiten dargestellt, dann hätte die zweite Variable eine größere Varianz als das erste Merkmal aufzuweisen. Diese Unterschiede manifestieren sich in den verschiedenen Dimensionen der Streuungsbereiche beider Variablen.

Durch eine Standardisierung werden diese Dimensionen größenmäßig angeglichen, mit einer Normierung auf den Merkmalsbereich erreicht man ihre vollkommene Übereinstimmung. Es können daher, wie schon oben erwähnt, durch eine Normierung beträchtliche Verformungen der Merkmalsräume hervorgerufen werden, die sich besonders auf die Ähnlichkeitsrelationen zwischen den Individuen auswirken.

Bei einer Normierung der Merkmale wird daher von vornherein angenommen, daß die Unterschiede in der Variabilität der Merkmale für die Klassifikation ohne Bedeutung sind. Daß diese Forderung nicht aufrechterhalten werden kann, soll folgendes Beispiel zeigen (vgl. SNEATH & SOKAL, 1973, 155): Es liegen verschiedene Populationen vor, die alle den selben Mittelwert haben, die Varianzen sind jedoch verschieden. Wenn man die Populationen anhand der Varianzen größenmäßig ordnet und dann standardisiert, so kommt mit abnehmender Größe der Varianzen diesen durch die Normierung immer mehr Gewicht zu. Dadurch wächst die Bedeutung der Varianzen bei ihrer Abnahme. Bei dem auch in der Praxis auftretenden Varianzwert von 0, d. h., wenn das Merkmal keine Variabilität zeigt, ist diese Variable für die Klassifizierung der Objektmenge jedoch zur Bedeutungslosigkeit degradiert, da ein nichtvariierendes Merkmal keine inneren Strukturen einer Objektmenge aufzeigen kann.

Dieser Widerspruch tritt jedoch nicht auf, wenn man mehrere Objektmengen vergleicht. Folgende Überlegung möge dies verdeutlichen: Stehen die Populationen mit abnehmender Varianz bei gleichen Mittelwerten in einer zeitlichen Abfolge, so zeigt sich das Bild der Auswirkung einer stabilisierenden Selektion, welche die Variationsbreiten bei konstantem Mittelwert verringert. Das kann dazu führen, daß die Merkmalsausprägungen nur mehr den Mittelwert als eine konstante Größe annehmen können, er wäre dann genetisch fixiert. Dabei hat aber die Variable eine Wandlung im Merkmalstyp durchgemacht, aus einem metrischen wurde ein klassifikatorisches Merkmal. Innerhalb der Objektmenge ging die Varianz des Merkmals verloren, bei einer anderen Objektmenge könnte sie noch vorhanden sein. Es besteht aber auch die Möglichkeit, daß in der zweiten Objektmenge ein anderer Merkmalswert fixiert wurde oder das Merkmal überhaupt verlorenging. Vom Standpunkt der Klassifikationsmöglichkeit zwischen den Objektmengen kommt den klassifikatorischen Merkmalen größeres Gewicht zu. Dies bedeutet, daß für eine Klassifikation innerhalb einer Objektmenge sich nichtändernde Merkmale keinen Wert haben, jedoch zwischen Objektmengen mit unterschiedlichen Merkmalsausprägungen große Bedeutung gewinnen.

Als Beispiel möge die Zahl der Halswirbel bei den Wirbeltieren dienen. Die Anzahl der Wirbel schwankt innerhalb der verschiedensten Klassen der Vertebraten, bei den Säugetieren ist aber eine Fixierung auf sieben Halswirbel eingetreten. Wollte man die einzelnen Ordnungen innerhalb der Säugetiere charakterisieren, ist dieses Merkmal zur Bestimmung von Gruppen ungeeignet, da bei allen Ordnungen die Konstanz der Anzahl der Halswirbel gewährleistet ist. Klassifiziert man jedoch die Wirbeltiere, so ist die Zahl der Halswirbel ein wesentliches taxonomisches Merkmal, da man durch sie die Säugetiere eindeutig von den anderen Klassen trennen kann.

Ein anderes Normierungsproblem stellt sich durch die Unterschiede in den Merkmalsarten, auch wenn sie in Skalen gleichen Typs zu transformieren sind. Wie schon erwähnt, sind im Euklidischen Raum die Distanzen als Ähnlichkeitsmaße besonders ausgezeichnet. Im multivariaten Merkmalsraum werden bei ihren Bestimmungen die Quadrate der Differenzen zwischen den einzelnen Merkmalswerten summiert und dann die Quadratwurzel gezogen (vgl. Formel 3). Bei dieser Vorgangsweise muß jedoch vorausgesetzt werden, daß den Merkmalen äquivalente Bedeutung zukommt. Wenn in einer multivariaten Analyse beispielsweise Körpergröße, Alter und Gewicht als Merkmale eingehen, werden sie zwar alle in einer metrischen Skala erfaßt, bei der Bestimmung der Distanzen zwischen den Individuen erfolgt dann aber eine Summierung von Differenzen unterschiedlicher physikalischer Größen. Dazu kommt noch, daß die Maßeinheiten der Skalen nicht vergleichbar sind. Wenn man die Merkmalswerte normiert, werden zwar die Wertebereiche der Variablen egalisiert, die Summierung von Distanzen unterschiedlicher Qualität bleibt jedoch erhalten. Die Annahme, daß gleiche Differenzen, die auf normierten Merkmalswerten beruhen, bei den unterschiedlichen Merkmalsarten die selbe klassifikatorische Bedeutung haben, läßt sich nicht begründen. Deshalb werden von JARDINE & SIBSON, 1971, 32, die Distanzen als Erfassung der Ähnlichkeitsrelationen zwischen Objekten abgelehnt und ein Informationsmaß bevorzugt, das sowohl skalen- als auch von den Merkmalstypen unabhängig ist (vgl. GOODALL, 1966; ORLÓCI, 1970, usw.). Die in der vorliegenden Arbeit vorgestellte Art der Standardisierung berücksichtigt jedoch auch jene Informationen, die in den Häufigkeitsverteilungen der einzelnen Merkmale, den Randverteilungen des multidimensionalen Merkmalsraums, stecken. Es besteht dann kein Grund, die Distanzen als Schätzung der Ähnlichkeiten von einer multivariaten Klassifikationsmethode auszuschließen.

Ein weiteres Problem stellt sich bei der Standardisierung, die, wie schon vorher erwähnt, die geläufigste Form der Normierung ist. Vom klassifikatorischen Standpunkt betrachtet sollte eine Standardisierung nur an symmetrisch verteilten Populationen durchgeführt werden. Wenn die Merkmalswerte einer gut strukturierten, d. h. in Gruppen aufteilbaren Objektmenge standardisiert werden, sollten die Summen der Häufigkeitsverteilungen der Klassen bei jeder Randverteilung normal- oder zumindest symmetrisch verteilt sein, d. h., durch die Summierung der einzelnen Häufigkeitsverteilungen sollte als Summe eine Normalverteilung entstehen. Nur in diesem Fall, wie in unserem Beispiel Abb. 1, ist eine Standardisierung sinnvoll. Wenn jedoch keine symmetrischen Randverteilungen vorliegen, führt deren Standardisierung zu starken,

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

über die durch die gleiche Gewichtung der Varianzen hervorgerufenen Verzerrungen des ursprünglichen Merkmalsraumes hinausgehenden Deformationen (vgl. Abb. 4). Diese zusätzlichen Verzerrungen werden durch die Abweichungen der Verteilungen von einer symmetrischen Verteilung hervorgerufen und stellen bei einer Abschätzung der gesamten Deformation einen wesentlichen, jedoch nur schwer erfaßbaren Anteil dar.

Die Abb. 4 möge diese Überlegungen verdeutlichen. Nur bei normalverteilten Variablen, mögen sie sich in der Größe der Lage- und Streuungsparameter noch so sehr unterscheiden, bewirkt eine Standardi-

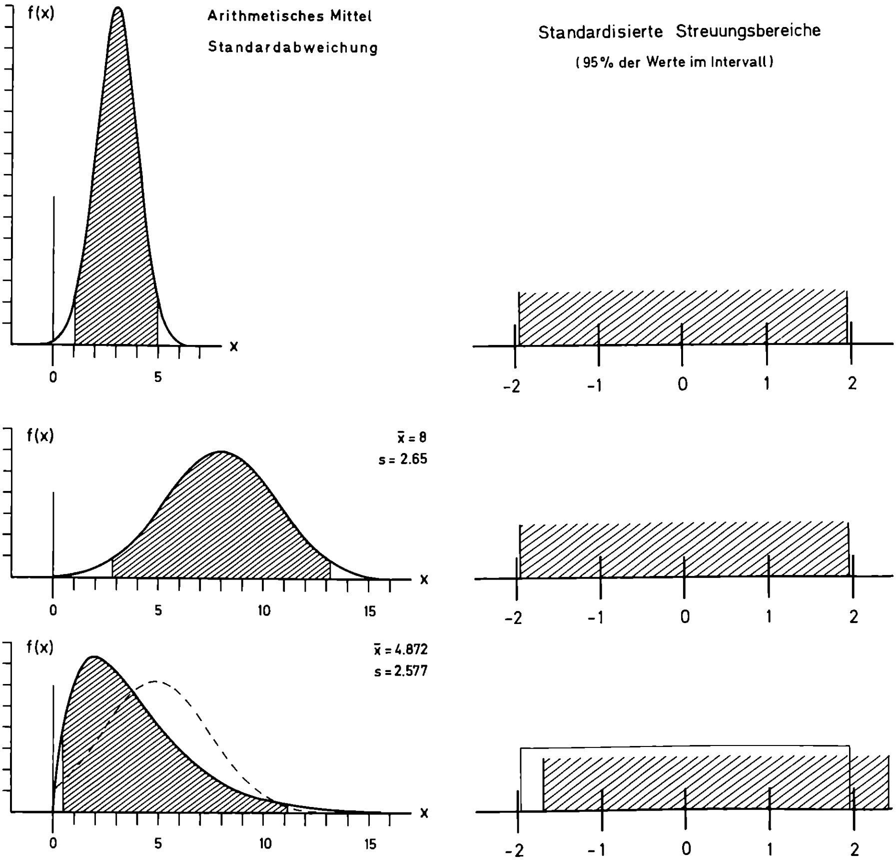


Abb. 4: Standardisierung nicht-normalverteilter Variablen und ihre Auswirkungen.
Die Standardisierung normalverteilter Variablen mit unterschiedlichen Verteilungsparametern (arithmetische Mittel und Standardabweichungen) erbringt übereinstimmende Streuungsbereiche (schraffierte Bereiche innerhalb der Intervallgrenzen im 1. und 2. Beispiel). Eine Standardisierung nicht-normalverteilter Variablen (3. Beispiel) ändert durch die Abweichung von der Normalverteilung (strichlierte Kurve) die Lage der normierten Streuungsbereiche. Das Ziel der Standardisierung, gleiche Lage und Wertebereiche, ist nicht mehr gewährleistet.

sierung eine gleiche Dimensionierung der Wertebereiche. In unserem Beispiel wurden zur Abschätzung der Streuungsbereiche die Intervallgrenzen genommen, zwischen denen 95 % der Individuen zu liegen kommen. Bei einer Standardisierung normalverteilter Variablen sind diese Grenzen in allen Verteilungen gleich (± 1,96). Wird aber eine asymmetrisch verteilte Variable standardisiert, so entstehen durch die Unterschiede in der Lage und Größe der Verteilungsparameter Differenzen zu einer Normalverteilung besonders in den Grenzen der Streuungsbereiche (vgl. Abb. 4, 3. Beispiel). Diese Abweichungen sind nun, da sie sehr stark von der jeweiligen nicht normalverteilten Häufigkeitsfunktion abhängen, nur schwer abzuschätzen, insbesondere wenn man in einer multivariaten Klassifikationsanalyse kritiklos alle Merkmale standardisiert und nicht die Form der Randverteilungen in Betracht zieht. Welche Vorteile deren Berücksichtigung bringt, mögen die folgenden Ausführungen zeigen.

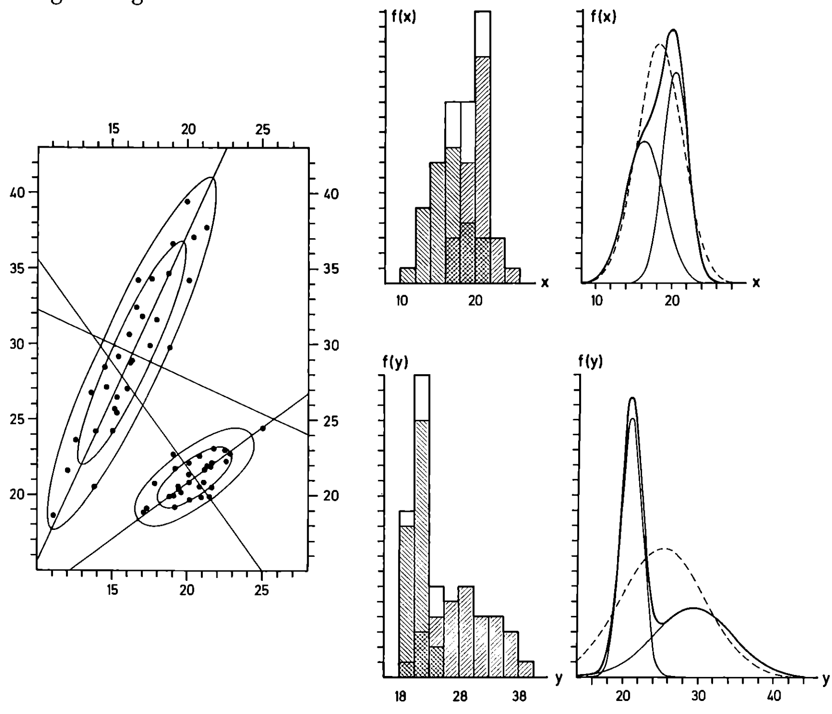


Abb. 5: Konfiguration von 60 Individuen im zweidimensionalen Euklidischen Merkmalsraum.
Wie in Abb. 1 sind die Verteilungsellipsen und Hauptachsen der beiden Klassen mit jeweils bivariaten Normalverteilungen dargestellt. Die Randverteilungen der Merkmale x und y sind als Häufigkeitsfunktionen in Form von Histogrammen (empirische Häufigkeitsfunktion) und Funktionsgraphen (theoretische Häufigkeitsfunktionen) dargestellt. Während die Häufigkeitsfunktion der Variablen x als Summe zweier Normalverteilungen wiederum einer Normalverteilung (strichlierte Kurve) ähnelt, weicht sie bei der Variablen y sehr stark von einer Normalverteilung ab.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Um die Vorstellung zu erleichtern, soll wiederum ein zweidimensionaler Merkmalsraum Verwendung finden. Im Streuungsdiagramm der Abb. 5 sind die Positionen der Individuen in einem von zwei Merkmalen aufgespannten Raum eingetragen. Es handelt sich hier, wie auch schon im ersten Beispiel (Abb. 1), um hypothetische Merkmalswerte, die durch einen statistischen Zufallsgenerator gewonnen wurden.

Betrachtet man die empirischen Häufigkeitsverteilungen der beiden Merkmale (Histogramme), so ist auf einen Blick zu erkennen, daß die Variable y nicht normalverteilt ist. Die Randverteilung weist eine extreme Schiefe auf, wobei der häufigste Wert (Modalwert) sich im unteren Bereich der Merkmalsskala bewegt. Aus den oben genannten Gründen würde hier eine Standardisierung starke Verzerrungen bewirken, die sich nicht nur durch die Angleichung der Wertebereiche erklären lassen, sondern auch noch auf den Abweichungen dieser Verteilungsform von einer Normalverteilung beruhen.

Doch bevor hier auf diese Deformationen eingegangen wird, soll dieses Beispiel dem morphologisch versierten Leser anschaulich gemacht werden (vgl. Abb. 6). Es wurden aus den beiden Gruppen, die sich in Abb. 5 deutlich unterscheiden, insgesamt acht Individuen ausgewählt und mit den Buchstaben A bis H benannt. Die Merkmale kann man z. B. als Größenmaße von Gastropodengehäusen deuten, wobei die erste Variable die Breite und das zweite Merkmal die Höhe des Schneckengehäuses darstellen soll. Außerdem wurde die Annahme getroffen, daß stets das gleiche Wachstumsstadium vorliegt, was sich in der konstanten Zahl von acht Windungen bei gleicher Größe der Anfangswindung ausdrücken soll.

Anhand dieser Voraussetzungen lassen sich die beiden Gruppen folgendermaßen charakterisieren: Eine Klasse beinhaltet hoch-trochospirale Formen, die in den Höhenabmessungen stark variieren, die andere Gruppe setzt sich aus flach-trochospiralen Gehäusen zusammen, bei denen die Streuungen in der Breite etwas größer sind. Nun wurde aus jeder Gruppe ein durchschnittliches Exemplar (Individuen C und D) mit jeweils zwei mehr oder minder extremen Gehäusen der anderen Klasse verglichen (vgl. Abb. 6, Individuen A und B bzw. E und F). Da die Distanzen in diesem zweidimensionalen Merkmalsraum die Ähnlichkeiten repräsentieren, ergeben sich bei einer Feststellung von Ähnlichkeitsbeziehungen zwei Dreiecke (ABC und DEF), wobei klar ersichtlich ist, daß bei dieser Gruppenkonfiguration die Individuen innerhalb einer Klasse ähnlicher sind, d. h. geringere Distanzen aufzuweisen haben, als Individuen, die verschiedenen Klassen angehören. In der geometrischen Darstellung ergeben sich zwei Dreiecke, für die gilt

$$\overline{AB} < \overline{AC} \quad \text{und} \quad \overline{AB} < \overline{BC}$$

sowie

$$\overline{EF} < \overline{DE} \qquad \overline{EF} < \overline{DF}.$$

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

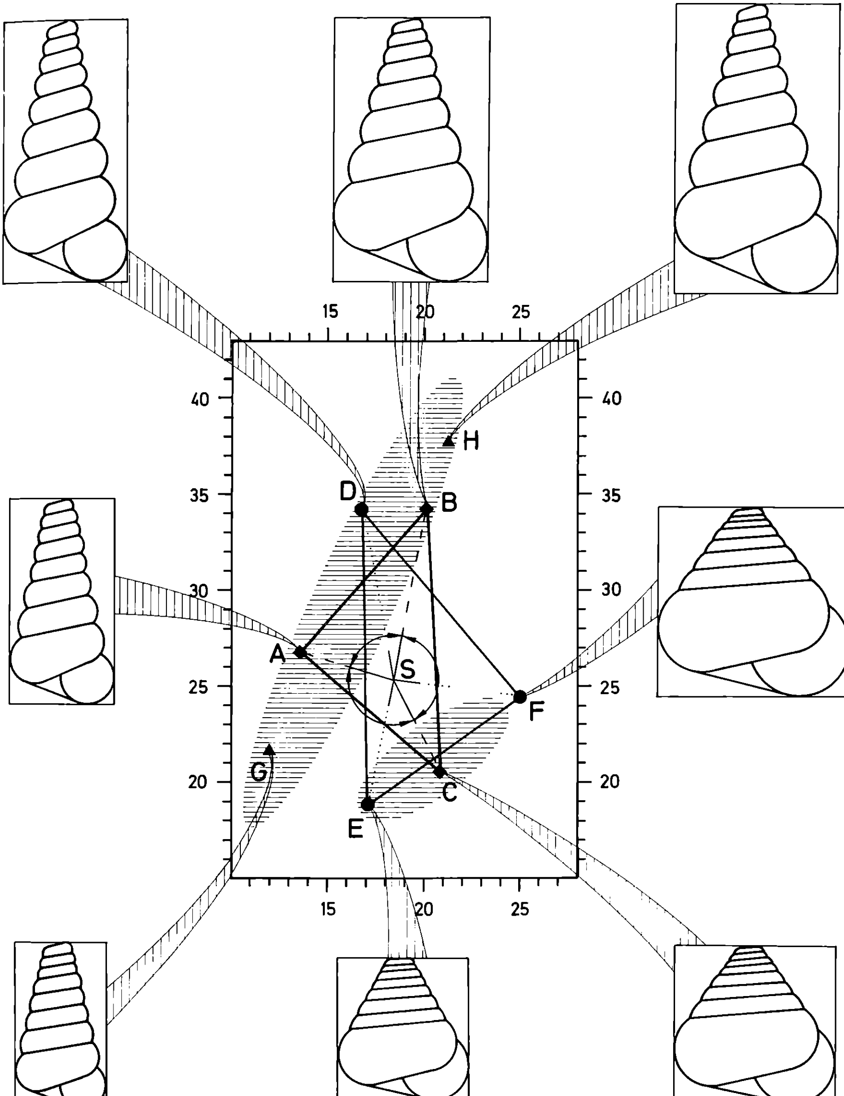


Abb. 6: Interpretation des Merkmalsraumes der Abb. 5 als Breite (Abszisse) und Höhe (Ordinate) trochospiral aufgerollter Gehäuse. Alle Gehäuse haben gleiche Größen der Anfangswindung und konstante Windungszahlen.
Im Diagramm sind die Streuungsbereiche der beiden Klassen, die 90 % der Individuen beinhalten, als schraffierte Flächen ausgeschieden. Die Buchstaben A bis H kennzeichnen einige markante Individuen, deren Gehäuseformen am Rand der Abbildung aufscheinen. Die Distanzen zwischen den Individuen A, B und C bzw. D, E und F kennzeichnen die Ähnlichkeitsbeziehungen zwischen diesen Individuen. Als weiteres Ähnlichkeitsmaß (Korrelationskoeffizient) kann der Winkel zwischen den Individuen vom Schwerpunkt der gesamten Verteilung (S) aus angesehen werden.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

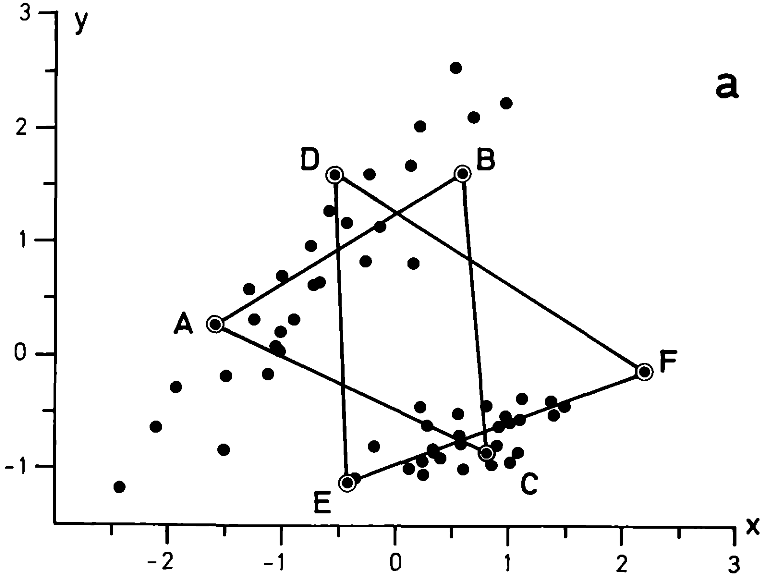
Die Euklidischen Distanzen zwischen den Punkten determinieren die Größe der Ähnlichkeit zwischen den Individuen.

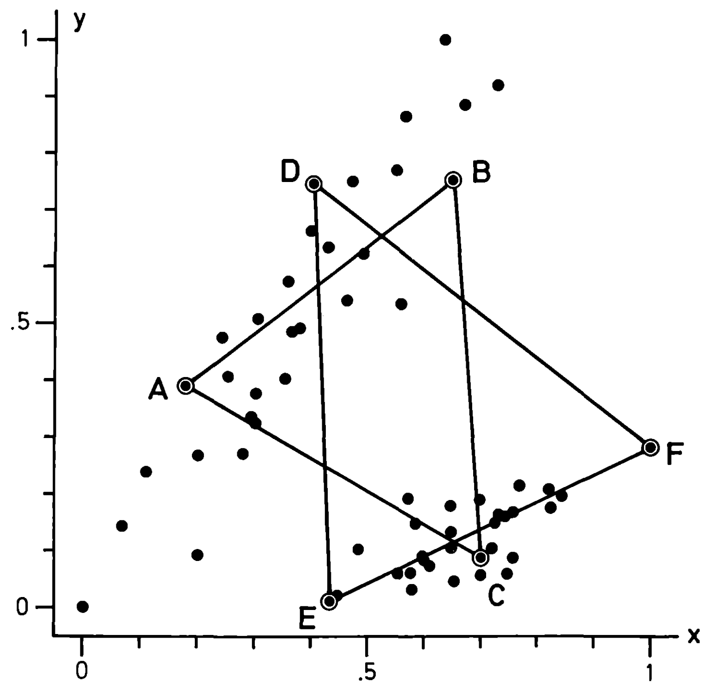
An dieser Stelle sei, um den Wert der Distanzen hervorzuheben, ein kurzer Hinweis auf Winkelmessungen als Ähnlichkeitsmaße gestattet. Der am häufigsten verwendete Koeffizient auf der Basis von Winkelbestimmungen ist der Produktmomentkoeffizient zwischen Objekten (vgl. BOYCE, 1969). In unserer Abb. 6 stellt er den Winkel zwischen zwei Individuen dar, der vom Schwerpunkt der gesamten Verteilung aus gemessen wird (Punkt S in Abb. 6). Es zeigt sich, daß die Positionen der Individuen im Merkmalsraum nicht durch Punkte, sondern durch Gerade, die die Individuen mit dem Schwerpunkt verbinden, determiniert sind. Dadurch läßt sich die Struktur des Merkmalsraumes nicht eindeutig bestimmen, wie es beispielsweise bei den Distanzen möglich ist.

Doch zurück zum Normierungsproblem. Wenn man in unserem Beispiel Abb. 5 eine Normierung durchführt, kommt es zu deutlichen Verzerrungen. Abb. 7 soll diese Deformationen zeigen. Es wurde eine Normierung auf den Merkmalsbereich (Abb. 7 b) einer Standardisierung (Abb. 7 a) gegenübergestellt. Die Normierung auf das Intervall [0,1] bewirkt, daß die Wertebereiche in ihren Ausdehnungen völlig gleich sind (vgl. Abb. 7 b). Vergleicht man die Ähnlichkeitsrelationen der bevorzugten Punkte A–F, so wird die Verzerrung gegenüber dem ursprünglichen Merkmalsraum (Abb. 6) deutlich sichtbar, die ursprüngliche Form der Dreiecke geht verloren.

Noch stärkere Deformationen treten bei der Standardisierung auf (Abb. 7 a), die dadurch hervorgerufen werden, daß die Randverteilung der zweiten Variablen keiner Normalverteilung entspricht. Der normierte Wertebereich des ersten Merkmals liegt, wie es bei einer Standardisierung normalverteilter Variablen zu fordern ist, symmetrisch um den 0-Punkt innerhalb der Intervallgrenzen von −2 bis + 2. Entgegengesetzt dazu ist der Wertebereich des zweiten Merkmals asymmetrisch um das arithmetische Mittel, dem 0-Punkt, gelagert (Intervall zwischen – 1,2 und + 2,5), wodurch extreme Verformungen in den Ähnlichkeitsrelationen hervorgerufen werden. Die Punkte C und D der Fremdpopulationen sind nun in der Ähnlichkeit den extremen Individuen A, B bzw. E, F deutlich nähergerückt. In diesem Fall gelten bei der Überführung der Individuen vom ursprünglichen in den standardisierten Merkmalsraum die oben erwähnten Monotoniebedingungen nicht mehr.

Abb. 7: Verzerrungen der Distanzen (= Ähnlichkeitsrelationen) des ursprünglichen Merkmalsraumes der Abb. 5 durch a) Standardisierung, b) Normierung auf den Merkmalsbereich.
Die Ähnlichkeitsbeziehungen im ursprünglichen Merkmalsraum zwischen jeweils drei Individuen (A – B – C und D – E – F) sind als Dreiecke in Abb. 6 gegeben. Beim Vergleich dieser ursprünglichen Konfiguration mit den Ähnlichkeitsrelationen in den normierten Merkmalsräumen zeigt die Standardisierung durch die nicht-normalverteilte Häufigkeit im Merkmal Gehäusehöhe (Variable y der Abb. 5) noch stärkere Deformationen der Dreiecke (= Ähnlichkeitsrelationen) als die Normierung auf den Merkmalsbereich.


a
y
x
3
2
1
0
-1
-2
-1
0
1
2
3
A
B
C
D
E
F


b
y
x
1
.5
0
0
.5
1
A
B
C
D
E
F

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Die Abb. 8 soll diese Verformungen des Merkmalsraumes noch verdeutlichen und gleichzeitig einen Hinweis auf den Informationsgehalt von Merkmalen liefern.

Werden bei den Gastropodengehäusen nur die Breiten und Höhen als taxonomische Merkmale betrachtet, vergleicht man zwei lineare Größen, die normal aufeinander stehen. In der geometrischen Darstellung ergeben diese beiden Maße Rechtecke, die nun das gesamte Gastropodengehäuse charakterisieren sollen. Daraus resultiert jedoch ein

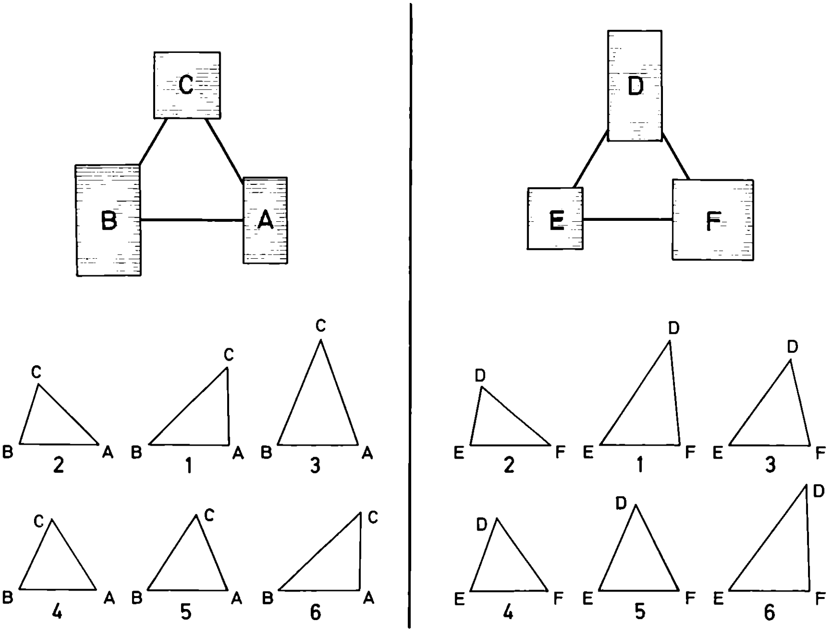


Abb. 8: Vergleich der Ähnlichkeitsrelationen als Distanzen zwischen den Individuen A, B und C bzw. D, E und F (vgl. Abb. 6) innerhalb verschiedener Merkmalsräume.
Um die Auswirkungen der unterschiedlichen Transformationen auf die Ähnlichkeitsrelationen zu verdeutlichen, wurden die Distanzen in den Abbildungen so normiert, daß die Strecken $\overline{AB}$ und $\overline{EF}$ stets einheitliche Größe aufweisen. Die beiden Merkmale eines Individuums, Gehäusehöhe und -breite, ergeben in der geometrischen Darstellung Rechtecke, deren Formen für die einzelnen Individuen im oberen Teil der Abbildung dargestellt sind.
Ähnlichkeitsrelationen als metrische Distanzen im
1. ursprünglichen Merkmalsraum,
2. Merkmalsraum mit unterschiedlichen Maßeinheiten (Gehäusebreite in cm, Gehäusehöhe in Inches),
3. (siehe unten),
4. einfach standardisierten Merkmalsraum,
5. auf den Merkmalsbereich normierten Merkmalsraum,
6. klassenspezifisch standardisierten Merkmalsraum.

Die Distanzen der mit der Ziffer 3 indizierten Dreiecke ergeben sich durch Transformationen der Korrelationskoeffizienten zwischen den Individuen im ursprünglichen Merkmalsraum.

erheblicher Informationsverlust, da bei einer solchen Analyse beispielsweise die Form und Anzahl der Windungen sowie das Wachstum keinerlei Berücksichtigung finden.

Die Rechtecke, durch die die einzelnen Individuen charakterisiert sind, werden in Abb. 8 (Individuen A, B, C und D, E, F) dargestellt. Die Distanzen als Ähnlichkeitsmaße erbringen die Relationen zwischen den Individuen in der Form von Dreiecken. Um die Unterschiede zwischen den einzelnen Konfigurationen zu verdeutlichen, wurde die Ähnlichkeit zwischen den Individuen einer Klasse (AB und EF) auf eine Einheitslänge normiert. Die mit der Ziffer 1 indizierten Dreieckspaare der Abb. 8 zeigen die Konfigurationen im ursprünglichen Merkmalsraum, die Dreiecke mit der Ziffer 2 sind dadurch bedingt, daß für die beiden Merkmale unterschiedliche Maßeinheiten verwendet wurden. So wurde in diesem Fall die zweite Variable (Gehäusehöhe) nicht in Zentimetern, sondern in Inches gemessen. Die Abweichungen von der ursprünglichen Konfiguration sind so stark, daß die Individuen der fremden Klasse zu denen der Stammgruppe größere Ähnlichkeit aufweisen als diese zueinander.

Die beiden Dreiecke mit der Ziffer 3 beruhen auf Ähnlichkeitsrelationen, die durch den Korrelationskoeffizienten bestimmt wurden. Dazu mußten die Werte der Koeffizienten in Distanzmaße übergeführt werden (vgl. BOCK, 1974, 77; SNEATH & SOKAL, 1973, 140). Die Ähnlichkeitsbeziehungen, die sich durch den Korrelationskoeffizienten ergeben, weichen stark von den Relationen ab, die durch die Euklidischen Distanzen repräsentiert werden. Da die Positionen der Individuen im Merkmalsraum bei den Winkelmessungen nicht eindeutig sind, können auch die Transformationen von Winkelmaßen in Distanzen die tatsächlichen Gegebenheiten nicht widerspiegeln.

Die Dreieckspaare mit den Ziffern 4 und 5 kennzeichnen die Ähnlichkeitsbeziehungen, wie sie durch eine Standardisierung (4) und eine Normierung auf den Merkmalsbereich (5) bewirkt werden. Es erweist sich, daß die Standardisierung durch die oben erwähnte zusätzliche Verzerrung des Merkmalsraumes die tatsächlichen Beziehungen noch schlechter widerspiegelt als die Normierung auf den Merkmalsbereich.

Aus diesen Beobachtungen kann folgender Schluß gezogen werden: Durch die bisher angewendete Normierung von Merkmalswerten gehen die tatsächlichen Ähnlichkeitsbeziehungen verloren. Der wesentliche Grund für diese starken Deformationen ist sicher in der Nichtberücksichtigung der Formen von Randverteilungen gelegen. Im folgenden Verfahren werden diese explizit berücksichtigt und vermeiden somit die gezeigten Nachteile anderer Verfahren.

In unserem Beispiel Abb. 5 ist die Objektmenge in zwei Klassen aufgeteilt, die ihrerseits in beiden Merkmalen Normalverteilung zeigen. Dies bedeutet jedoch, daß die Randverteilungen der Merkmale Summen von Häufigkeitsverteilungen darstellen, die aus jeweils zwei normalverteilten Komponenten resultieren. Da sich in den einzelnen Merkmalen die

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Komponenten sowohl in den Lage- als auch Streuungsparametern unterscheiden, kann man jeden Merkmalswert mit den Parametern der Komponenten standardisieren.

Durch diese Vorgangsweise erhöht sich aber der Merkmalsraum, da alle Merkmalswerte mit den Verteilungsparametern jeder Komponente zu standardisieren sind. Die Dimension des Merkmalsraumes hängt nun auch von der Zahl der Komponenten und nicht nur von der Anzahl der Merkmale ab. Sie ergibt sich aus dem Produkt der Zahl der Merkmale mit den Komponenten (m × h). Im vorgeführten Beispiel entsteht durch diese Form der Standardisierung ein vierdimensionaler Merkmalsraum, da jeder Merkmalswert mit den Verteilungsparametern sowohl der einen als auch der anderen Komponente standardisiert wird.

$$x^*_{ikl} = \frac{x_{ik} - \bar{x}_{kl}}{s_{kl}} \qquad (4)$$

i = 1,… n (n = Anzahl der Individuen);
k = 1,… m (m = Anzahl der Merkmale);
l = 1,.. h (h = Anzahl der Komponenten).

Durch die Erhöhung der Dimensionalität des Merkmalsraumes bei einer Standardisierung mit Komponenten erfolgt auch eine Änderung der Gleichung für die Euklidischen Distanzen:

$$D_{ij} = \left[\sum_{k=1}^{m} \sum_{l=1}^{h} (x_{ikl} - x_{jkl})^2\right]^{1/2} \qquad (5)$$

Da diese Form der Standardisierung auf die Verteilungsparameter der einzelnen Klassen zurückgreift, soll sie hier als „klassenspezifische Standardisierung“ bezeichnet und der einfachen Standardisierung gegenübergestellt werden.

Welche Vorteile bringt eine klassenspezifische Standardisierung? Hier kann wieder das Beispiel der Abb. 5 für die weiteren Überlegungen dienen. Bei der klassenspezifischen Standardisierung werden beispielsweise die Merkmalswerte der Variablen y nicht nur einmal, sondern zweimal standardisiert. Zuerst erfolgt eine Normierung mit dem Mittelwert und der Standardabweichung der Klasse mit den großen, schlanken Gehäusen. Anschließend wird mit den Verteilungsparametern der Gruppe mit den niedrigen, breiten Individuen standardisiert. Die beiden standardisierten Wertebereiche liegen nun nicht mehr symmetrisch um den Mittelwert 0 innerhalb der Intervallgrenzen −2 und + 2 angeordnet. Nur die Merkmalswerte der Individuen, die der Klasse angehören, mit deren Verteilungsparametern standardisiert wurde, streuen in diesen Bereichen. Alle Individuen, die außerhalb des Streuungsbereiches der Klasse zu liegen kommen, erweitern den standardisierten Merkmalsbereich. Den Grad der Erweiterung bestimmen die Distanzen dieser Individuen zum Mittelwert der standardisierten

Klasse. Als Maßeinheit für diese Distanzen gilt die Standardabweichung der Klasse, mit deren Verteilungszahlen die Normierung durchgeführt wurde. Somit hängen die Ausmaße der standardisierten Wertebereiche sehr stark von den Lage- und Streuungsparametern der einzelnen Klassen ab.

Als Beispiel möge das Merkmal y der Abb. 5 dienen. Die Häufigkeitsverteilung dieser Variablen läßt sich in zwei Klassen zerlegen, von denen die erste (schlanke Individuen) folgende Parameterwerte aufweist: Arithmetisches Mittel $\bar{y}_1 = 29{,}38$; Standardabweichung $s_y = 5{,}25$. Die Maßzahlen der zweiten Klasse (Formen mit dicken Gehäusen) lauten: $\bar{y}_2 = 21{,}18$; $s_y = 1{,}39$. Der mit den Parametern der 1. Klasse standardisierte Wertebereich des Merkmals Gehäusehöhe reicht von $-2{,}04$ bis $+ 1{,}92$, der Wertebereich des gleichen Merkmals, mit den Verteilungsmaßzahlen der 2. Klasse standardisiert, liegt im Intervall von $-1{,}81$ bis $+ 13{,}14$. Durch die relativ kleine Varianz der 2. Klasse und den kleineren Mittelwert bedingt liegt das eine Extrem des Wertebereichs mehr als 13 Standardabweichungen vom arithmetischen Mittel entfernt (vgl. Abb. 5, Häufigkeitsverteilung des Merkmals y).

Durch diese extreme Ausdehnung des Wertebereiches ergeben sich bei der Erfassung der multivariaten Distanzen wiederum Verformungen des Merkmalsraumes, die aber einer Verzerrung, wie sie durch die einfache Standardisierung hervorgerufen wird, entgegenwirken. Die Ähnlichkeit zwischen den Gehäusen A, B und C sowie D, E und F lassen sich, obwohl aus einem vierdimensionalen Merkmalsraum gewonnen, durch geeignete Rücktransformation als Dreiecke im zweidimensionalen Raum darstellen. In der Abb. 8 sind es die beiden mit der Ziffer 6 indizierten Dreiecke. Anhand ihrer Formen erkennt man deutlich, daß hier die besten Übereinstimmungen mit den Positionen im ursprünglichen Merkmalsraum gegeben sind.

Aus den empirischen Überlegungen folgt: Die klassenspezifische Standardisierung schafft analog zur einfachen Standardisierung skalenunabhängige Merkmalswerte, bewirkt im Gegensatz zu ihr aber keine Deformation der Strukturen des ursprünglichen Merkmalsraumes, sobald dieser in gleichen Maßeinheiten erfaßt wurde.

Informationsgehalt von Merkmalen

Im vorigen Kapitel wurde gezeigt, daß bei der Normierung von Merkmalswerten die klassenspezifische Standardisierung die beste Repräsentation der Punktekonfiguration einer strukturierten, d. h. in Klassen geteilten Objektmenge bringt. In der Abb. 5 wurden zwei Merkmale als Achsen im Euklidischen Raum dargestellt. Durch diese Form der Merkmalskombination wird zwar eine simultane Betrachtung der Variablen ermöglicht, die Zusammenhänge zwischen den Merkmalen bleiben jedoch unberücksichtigt. Da es sich um quantitative Variablen handelt, lassen sich ihre Relationen in einfacher Form durch den Quotienten ausdrücken (vgl. Abb. 9). Durch diese Vorgangsweise reduziert sich der Merkmalsraum auf eine Dimension, was aber notgedrungen zu einem Informationsverlust führen muß.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Wie läßt sich der Quotient zwischen den Merkmalswerten eines Individuums im zweidimensionalen Merkmalsraum darstellen? Seine Definition a = y/x kann auf die Funktionsgleichung y = ax zurückgeführt werden. Der Funktionsgraph dieser Gleichung entspricht einer Geraden, die den Ursprung des Koordinatensystems mit dem Punkt (x, y) im Merkmalsraum verbindet. Ihr Anstieg (Parameter a) läßt sich als Winkelfunktion (tan α) ausdrücken. Der Winkel α wird von dieser Geraden und der Abszisse des Koordinatensystems eingeschlossen.

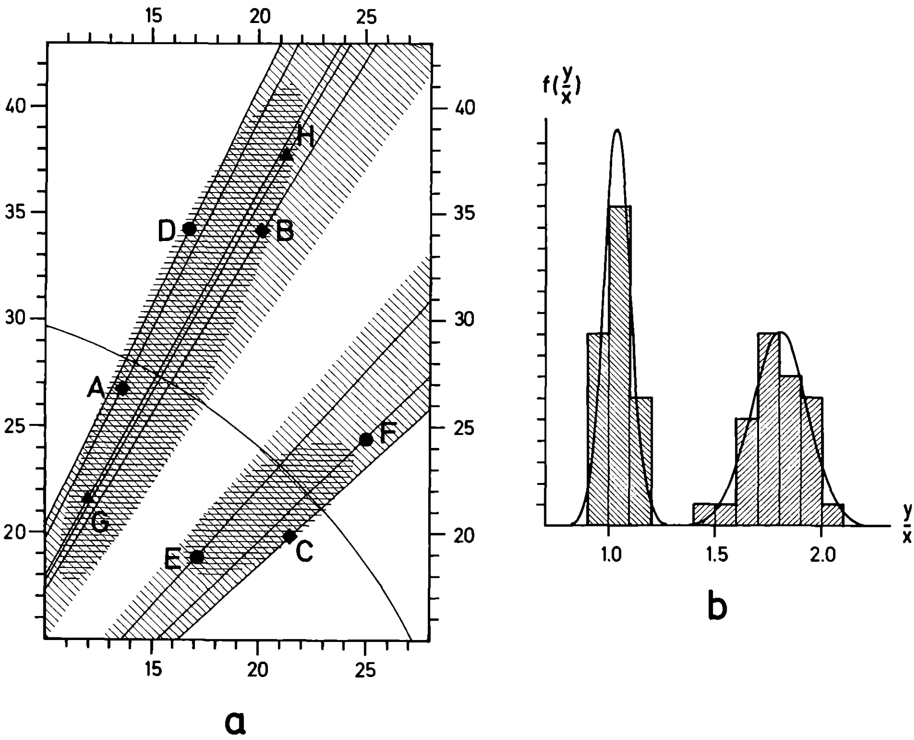


Abb. 9: a) Darstellung der Quotienten aus den beiden Merkmalen x und y der Abb. 5 im zweidimensionalen Merkmalsraum.
In dieser Form sind die Quotienten als Anstiege von Geraden repräsentiert, die den Ursprung des Koordinatensystems mit den einzelnen Punkten verbinden. Unterschiede zwischen den Individuen (Kosinus-Koeffizienten) lassen sich als Abschnitte auf einem Kreisbogen darstellen, deren Grenzen durch die Schnittpunkte der Geraden mit dem Kreisbogen gegeben sind. Das Zentrum des Kreises liegt im Ursprung des Koordinatensystems.
Wie in Abb. 6 sind die Streuungsbereiche der beiden Klassen durch schraffierte Ellipsen dargestellt, die schräge Schraffur kennzeichnet die Streuungsbereiche der Quotienten.
b) Empirische und theoretische Häufigkeitsfunktionen der Quotienten y/x der Objektmenge von Abb. 5.
Die Trennung der beiden Klassen ist durch die unterschiedlichen Verteilungen ohne Überlappung der Streuungsbereiche verdeutlicht.

Den Unterschied zwischen zwei Individuen im Merkmalsraum (als Ähnlichkeit definierbar) kann man auch mit den Größendifferenzen der beiden Winkel erfassen. Als Maß ist der Tangens jedoch ungeeignet, da er zwischen den Werten 0 und ∞ schwankt und größere Abweichungen bis zu einem extremen Ausmaß gewichten würde. Aus diesem Grund wird als Ähnlichkeitsmaß meist der Kosinus verwendet. Er ähnelt in seinem Wertebereich dem Korrelationskoeffizienten, wobei die Werte 1 vollkommene Übereinstimmung in den Merkmalswerten, 0 aber totale Diskrepanz der Individuen bedeuten. Der Kosinus-Koeffizient als Ähnlichkeitsmaß zwischen zwei Objekten läßt sich nicht nur im zwei-, sondern auch im multidimensionalen Merkmalsraum berechnen, wobei sein Wertebereich von 0 bis 1 erhalten bleibt. Die Kosinus-Koeffizienten sind daher als multidimensionale Differenzen von Merkmalsquotienten zu definieren.

Schon im zweidimensionalen Raum ist die Bestimmung von Quotienten problematisch, wie dies beispielsweise ZORN, 1972, deutlich vor Augen führt. Werden Quotienten berechnet, so wird von vornherein angenommen, daß innerhalb der Klassen die Merkmalsbeziehungen linear sind und diese Geraden stets durch den Ursprung gehen. In biologischen Populationen sind jedoch meist solche Zusammenhänge nicht gegeben, insbesondere da das Wachstum eine wesentliche Komponente des Zusammenhanges von Merkmalen ist. Ein nichtlinearer Zusammenhang (allometrisches Wachstum) ist hier die Regel, isometrisches (lineares) Wachstum stellt die Ausnahme dar. Durch das allometrische Wachstum ändern sich die Streuungsbereiche der Quotienten mit dem Alter kontinuierlich, während sie beim isometrischen Wachstum konstant bleiben.

Im Beispiel Abb. 6 wurde stets dasselbe Wachstumsstadium angenommen, wodurch die Einflüsse des Wachstums bei der Quotientenbildung ausgeschlossen wurden. Die Häufigkeitsverteilung der Höhe/Breite-Quotienten zeigt eine deutliche Trennung zweier normalverteilter Klassen (vgl. Abb. 9 b), zwischen den beiden Verteilungen liegt ein breiter Zwischenraum. Bei einer Konfiguration der Individuen im Merkmalsraum, wie sie in Abb. 5 gegeben ist, wäre die Trennung der Klassen mit Hilfe der Quotienten ideal. Sie ist auf alle Fälle deutlicher als in den Häufigkeitsverteilungen der einzelnen Merkmale (vgl. Histogramme der Abb. 5 mit der Abb. 9 b).

Eine optimale Trennung der Klassen einer strukturierten Objektmenge mit Hilfe der Quotienten kann jedoch nur dann erfolgen, wenn die Merkmale positiv korreliert sind, d. h., mit steigendem Wert des einen erhöht sich auch der Wert des anderen Merkmals. Liegt eine negative Korrelation vor, in unserem Beispiel würde sie durch eine Spiegelung um die Ordinate hervorgerufen werden, so umfassen die Klassen in den Quotienten ungefähr die selben Wertebereiche, und eine Trennung anhand der Häufigkeitsverteilung kann nicht mehr durchgeführt werden.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Will man ein multivariates Klassifikationsverfahren anwenden, das auf Ähnlichkeitsbestimmungen mittels Kosinus-Koeffizienten beruht, muß eine Überprüfung der positiven Korrelation von Merkmalen vorliegen. Auch wenn alle Korrelationsberechnungen zwischen jedem Merkmalspaar durchgeführt wurden, hat man damit noch nicht die Einflüsse erfaßt, die die Korrelationen untereinander haben. Dies könnte man durch Berechnung von partiellen Korrelationen abschätzen und ausschalten. Der Rechenaufwand würde dadurch jedoch immer größer, komplizierter und in der Folge nicht mehr rentabel, insbesondere da die Kosinus-Koeffizienten noch weitere Schwierigkeiten bereiten.

Wie bei jedem Winkelmaß – so auch bei den Quotienten und Kosinus-Koeffizienten – sind die Objekte nur in ihrer Richtung zum Ursprung des Merkmalsraumes festgelegt. Ihre exakten Positionen sind damit jedoch keineswegs fixiert, sie können entlang dieser Geraden an jeder Stelle gelegen sein. Damit gehen jedoch bei einem Vergleich von Individuen die Unterschiede in den absoluten Merkmalsgrößen verloren.

Nimmt man in den Abb. 6 und 9a die Individuen B, G und H, die alle der Gruppe mit den schlanken Gehäusen angehören, und betrachtet ihre Ähnlichkeitsrelationen auf der Basis von Quotienten, so können diese Beziehungen als Vergleiche von Abschnitten eines Kreisbogens, deren Grenzen die Schnittpunkte der Geraden sind, gedeutet werden (vgl. Abb. 9 a). In der Abb. 14/2 sind die Relationen sofort zu erkennen. Da mit den Quotienten nur die Höhe-Breite-Relationen und nicht die absoluten Größen erfaßt werden, sind die Individuen G und H ähnlicher als B und H bzw. B und G. Für den morphologisch geschulten Beobachter steht jedoch außer Zweifel, daß B und G einander wesentlich ähnlicher sind, wie es sich auch in der Konfiguration des ursprünglichen Merkmalsraumes (vgl. Abb. 6) ausdrückt.

Da die Positionen der Individuen im Merkmalsraum durch die Quotienten nicht eindeutig festgelegt sind, gehen innerhalb einer Gruppe die ursprünglichen Ähnlichkeitsrelationen verloren. Zwischen den Klassen sind Quotienten als Ähnlichkeitsmaße nur dann anzuwenden, wenn a priori eine günstige Konstellation mit einer zusätzlichen positiven Merkmalskorrelation gegeben ist. Ob eine solche Konstellation vorliegt, kann nur nach einem Klassifikationsvorgang erfaßt werden, so daß die Quotienten bzw. die Kosinus-Koeffizienten als Ähnlichkeitsmaße für eine Klassifikationsanalyse nur bei günstigsten Voraussetzungen brauchbare Ergebnisse liefern können.

Im vorigen Kapitel wurde bereits mehrmals auf den Informationsverlust hingewiesen, der durch die Wahl von Höhe und Breite als charakteristische Gehäusemerkmale von Gastropoden hervorgerufen wird. Nimmt man diese beiden „charakteristischen" Merkmale und versucht anhand dieser die Gastropodengehäuse zu rekonstruieren, so wird man mit den besten Bemühungen für jedes Individuum als geometrischen Körper nur Rechtecke gewinnen, die die Gestalt der Gehäuse keineswegs zu repräsentieren vermögen (vgl. Abb. 8, oberer

Teil). Wie sich dies in einer numerischen Klassifikationsmethode auswirken kann, soll der Vergleich der extremen Individuen B und G (Klasse mit den schlanken Gehäusen) mit einem mittleren Gehäuse der Fremdpopulation (Individuum C mit einem gedrungenen Gehäuse, vgl. Abb. 6) zeigen. Zieht man die Höhe und Breite – zwar kombiniert, aber dennoch als einzelne Merkmale betrachtet – als Variable heran und bestimmt die Ähnlichkeiten, so ergeben sich zwischen den Individuen einer Population (B und G) geringere Ähnlichkeiten (größere Distanzen) als zwischen den Individuen der unterschiedlichen Klassen (B zu C und C zu G; vgl. Abb. 15/1). Das entspricht aber keineswegs dem Empfinden eines Morphologen, der die drei Gehäuse vergleicht. Er würde auf alle Fälle die beiden Individuen B und G als ähnlicher bezeichnen und das Individuum C separieren. Eine Ähnlichkeitsrelation anhand der Quotienten (vgl. Abb. 15/2) entspricht zwar eher den Empfindungen des Morphologen, der C von den beiden Formen B und G deutlich trennen würde, doch, wie schon erwähnt, ist diese Trennung durch die günstige Konfiguration der Klassen im Merkmalsraum bedingt. Innerhalb der Gruppen stehen die auf Quotienten basierenden Ähnlichkeitsrelationen im starken Kontrast zur Empfindung des Morphologen (vgl. Abb. 14/2).

Somit zeigt es sich, daß die Höhen- und Breitenmaße zur Charakterisierung der Gehäuseformen nicht ausreichen. Darum muß hier auf andere, informationsträchtigere Merkmale zurückgegriffen werden. Wie bereits erwähnt, wurden als Beispiele Gastropodengehäuse herangezogen, die alle dasselbe Wachstumsstadium aufweisen. Durch die Arbeiten theoretischer Morphologen wurde gezeigt, daß sich trochospiral aufgerollte Gehäuse, wenn man sie als geometrische Formen im dreidimensionalen Raum mit zylindrischen Koordinaten definiert, fast vollständig durch vier Gleichungsparameter chrakterisieren lassen (vgl. Raup, 1966). Folgende Maßzahlen bestimmen die Gestalt (vgl. Abb. 10):

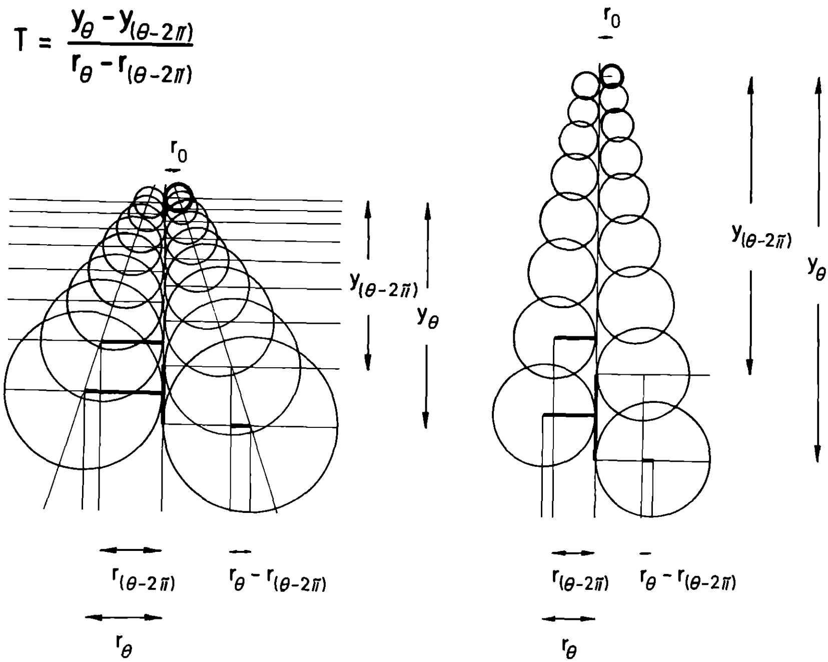
1. Die Form der Anfangswindung (Initialfigur): Sie wurde in unserem Beispiel als Kreis angenommen, der Radius (r_0) hat bei allen Individuen eine konstante Größe.
2. Die Position der Windungsfigur zur Aufrollungsachse (Parameter D): Auch dieser Parameter wurde für jedes Individuum konstant gehalten, wobei die Windungsquerschnitte die Aufrollungsachse berühren.
3. Die Expansionsrate des Zentrums der Anfangswindung (Parameter W): Der Parameter charakterisiert die stetige Zunahme der Distanz des Schwerpunktes der Windungsfigur zur Aufrollungsachse bei jeder Windung.
4. Die Versetzungsrate (Parameter T) stellt die Expansionsrate in Relation zur Bewegung entlang der Aufrollungsachse.

Wie schon erwähnt, haben konstante Werte bei einer Klassifizierung innerhalb einer Objektmenge keinerlei Bedeutung. Deswegen können auch in unserem Beispiel die beiden ersten Parameter (r_0 und D) bei den folgenden Überlegungen außer acht gelassen werden. Sowohl die Expansions- als auch die Versetzungsrate lassen sich bei trochospiral

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

$$W = \frac{r_\theta}{r_{(\theta-2\pi)}}$$

$$T = \frac{y_\theta - y_{(\theta-2\pi)}}{r_\theta - r_{(\theta-2\pi)}}$$



r_0 = konstant

D = konstant = 1/2

1.279 = W = 1.214

2.969 = T = 8.477

Abb. 10: Ermittlung der Gehäuseparameter von trochospiral aufgerollten Gehäusen.
Die Gehäuseform wird geometrisch als Rotation eines zweidimensionalen geometrischen Körpers (Initialfigur) um eine Achse bei stetiger Vergrößerung gedeutet. Diese Vergrößerung läßt sich mit folgenden Parametern mathematisch erklären (vgl. RAUP, 1966):
D = Position des geometrischen Körpers zur Aufrollungsachse.
W = Expansionsrate = Vergrößerungsrate.
T = Versetzungsrate entlang der Aufrollungsachse.
Die geometrische Konstruktion der Individuen F und D (siehe Abb. 6) mit gleichen Initialfiguren (r_0) und Positionen zur Aufrollungsachse bei unterschiedlicher Expansions- und Versetzungsrate ist dargestellt.

aufgerollten Gehäusen relativ leicht ermitteln (vgl. Formeln der Abb. 10). Sie besitzen gegenüber den Höhe- und Breitemaßen größeren Wert, da sie als Funktionsparameter vom Wachstum unabhängig sind. Unter diesen Voraussetzungen (konstantes r_0 und D, variables W und T) lassen sich die Gehäuse in ihrer Form vollständig durch die beiden Parameter W (= Expansionsrate) und T (= Versetzungsrate) charakterisieren. Bei einer Rückführung der Parameterwerte in geometrische Figuren ergeben sich die Formen von Abb. 6 und Abb. 12–15. Daraus läßt sich die gute Charakterisierung der trochospiralen Gehäuse durch die beiden Parameter W und T erkennen.

Welche Folgerungen ergeben sich, wenn man die beiden Parameter W und T als Merkmale betrachtet? Da es sich um abgeleitete Merkmale (= indirekte Messungen) handelt – sie wurden aus direkten Messungen gewonnen –, spielen die Maßeinheiten keine Rolle. Die absoluten Werte auf der Zahlengeraden, wie sie durch die Merkmale geliefert werden, variieren dennoch beträchtlich: Umfaßt der Wertebereich des Merkmals Expansionsrate das Intervall 1,155–1,279 mit einer Intervallbreite von 0,124, streuen die Werte für die Versetzungsrate im Bereich 2,791–8,499 (Intervallbreite: 5,708).

Anhand der beiden neuen Merkmale kann eine Klassifikation der Gehäuse erfolgen, die eine bessere Trennung der Klassen ergeben müßte, da diese Parameter wesentlich mehr Information über die Gehäusegestalt liefern als die bisherigen Höhen- und Breitenmaße.

Hier beginnen aber alle Probleme, die in den vorigen Kapiteln eingehend behandelt wurden, im vollen Umfang einzusetzen. Beispielsweise wurde gefordert, daß die Merkmale kombiniert zu betrachten sind, was bei der Zahl der vorliegenden Merkmale durch ein Streuungsdiagramm gewährleistet ist. Abb. 11 a zeigt jedoch den durch die ungleichen Wertebereiche der Variablen hervorgerufenen Informationsverlust im ursprünglichen Merkmalsraum, der von der Expansions- und Versetzungsrate aufgespannt wird. Durch die große Ausdehnung ihres Wertebereiches gewinnt die Versetzungsrate gegenüber der Expansionsrate großes Gewicht. Im vorliegenden Fall wäre dies nicht so sehr von Nachteil, da bei diesem Merkmal die Trennung der beiden Klassen in ähnlichem Maß wie bei den Höhe/Breite-Quotienten gewährleistet ist. Durch die extrem geringe Dimension des Wertebereiches der Variablen W ist aber das Streuungsdiagramm Abb. 11 a nur zu einer anderen Abbildungsform der Häufigkeitsverteilung der Variablen T degradiert (vgl. Abb. 11 c). Bei dieser Konfiguration geht, wollte man Ähnlichkeitsbeziehungen zwischen den Individuen ermitteln, die im Merkmal W enthaltene Information vollständig verloren. Darum muß eine Angleichung der Wertebereiche gefordert werden. Hier wurde in anderer Form als bei einer Normierung vorgegangen. Man muß sich nicht unbedingt an festgelegte Transformationen halten, sondern kann individuell die Merkmalsräume angleichen, wie es fast stets bei einer Darstellung von zwei Merkmalen in einem Streuungsdiagramm geschieht. Bei der

Abb. 11 b geschah dies, indem der Wertebereich der Versetzungsrate auf 1/100 der ursprünglichen Größe verringert wurde. Das überraschende Ergebnis dieser Transformation ist die deutliche Differenzierung in den Wertebereichen der beiden Klassen auch im Merkmal Expansionsrate, die in der ursprünglichen Konfiguration nicht zu erkennen war. Die niedrigtrochospiralen Gehäuse liegen in den oberen Bereichen der Merkmalsskala, hoch trochospirale Gehäuse sind eher in den unteren Bereichen

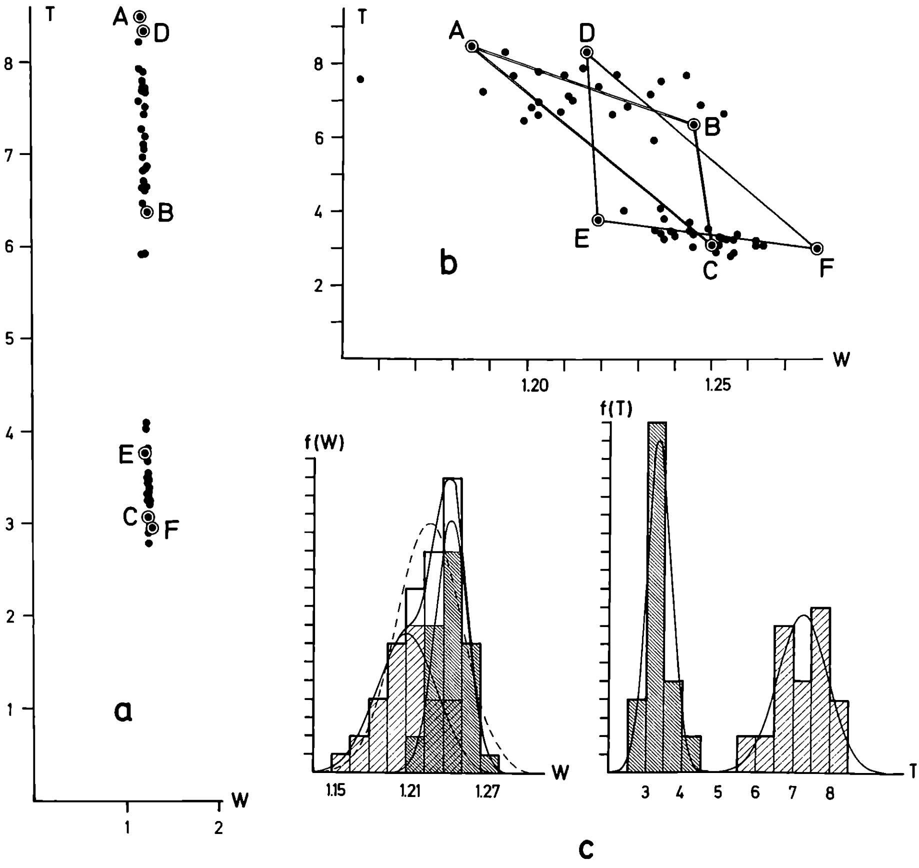


Abb. 11: a) Konfiguration der 60 Individuen von Abb. 5 in einem zweidimensionalen Merkmalsraum, der von den Variablen Expansionsrate (= W) und Versetzungsrate (= T) aufgespannt wird.
b) Transformation des Wertebereiches des Merkmals Versetzungsrate (= T) auf 1/100 der Größe im ursprünglichen Merkmalsraum (Abb. 11a). Diese Transformation führt zu starken Verzerrungen in den Ähnlichkeitsbeziehungen zwischen den Individuen.
c) Darstellung der empirischen und theoretischen Häufigkeitsverteilungen der Variablen Expansions- und Versetzungsrate in der Form von Histogrammen und Funktionskurven. Bei der Versetzungsrate T ist die Trennung der beiden Klassen in ähnlichem Maß wie bei den Quotienten gegeben (vgl. Abb. 9b). Im Gegensatz dazu nähert sich die Expansionsrate W in ihrer Summenfunktion einer Normalverteilung (strichlierte Kurve).

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

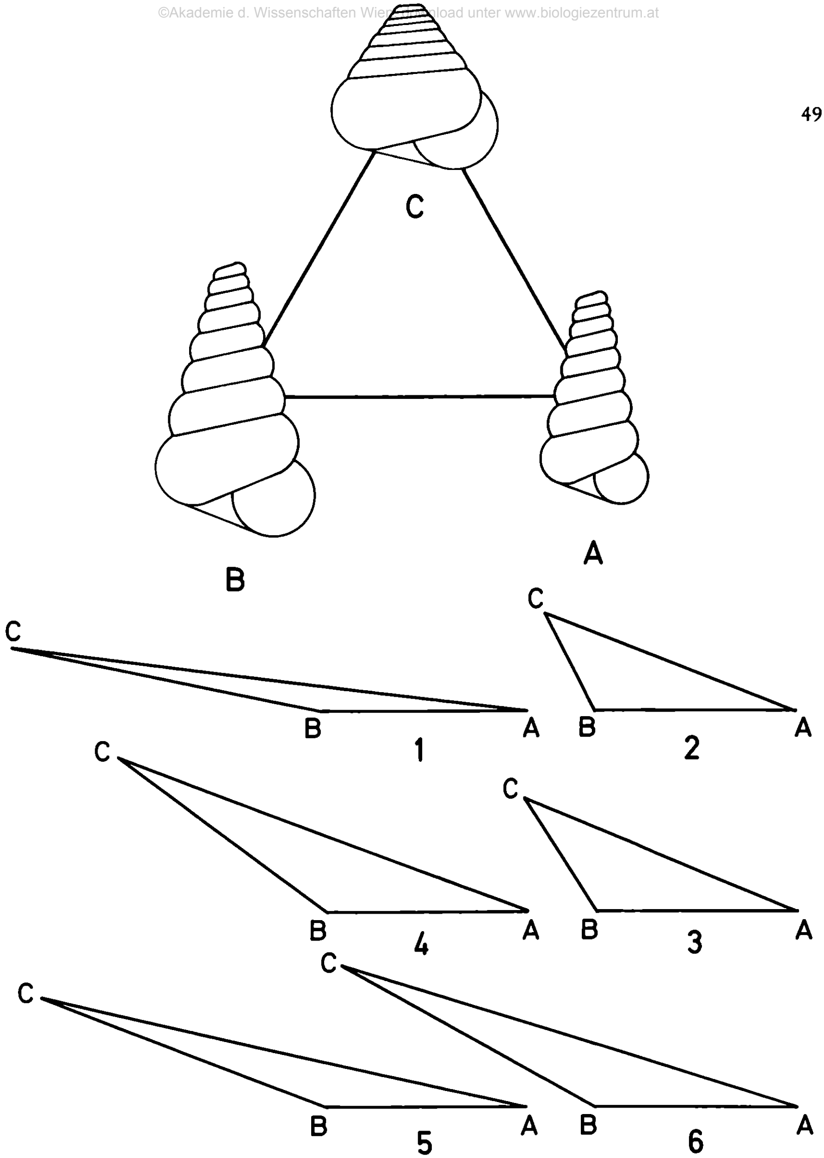


Abb. 12: Vergleiche der Ähnlichkeitsrelationen zwischen den Individuen A, B und C in den unterschiedlich transformierten Merkmalsräumen, die von den Variablen Expansionsrate (W) und Versetzungsrate (T) aufgespannt werden.

Die beiden Merkmale W und T bestimmen die Form der Gehäuse, wie sie im oberen Teil der Abbildung dargestellt sind. Die Ähnlichkeitsrelationen sind so normiert, daß die Distanzen $\overline{AB}$ stets die gleiche Länge aufweisen.

Ähnlichkeitsrelationen als metrische Distanzen im

1. ursprünglichen Merkmalsraum (Abb. 11a),
2. transformierten Merkmalsraum (Abb. 11b),
3. einfach standardisierten Merkmalsraum,
4. klassenspezifisch standardisierten Merkmalsraum,
5. komponentenspezifisch standardisierten Merkmalsraum mit ungleicher Komponentenzahl in den Merkmalen,
6. komponentenspezifisch standardisierten Merkmalsraum mit gleicher Komponentenzahl in den Merkmalen.

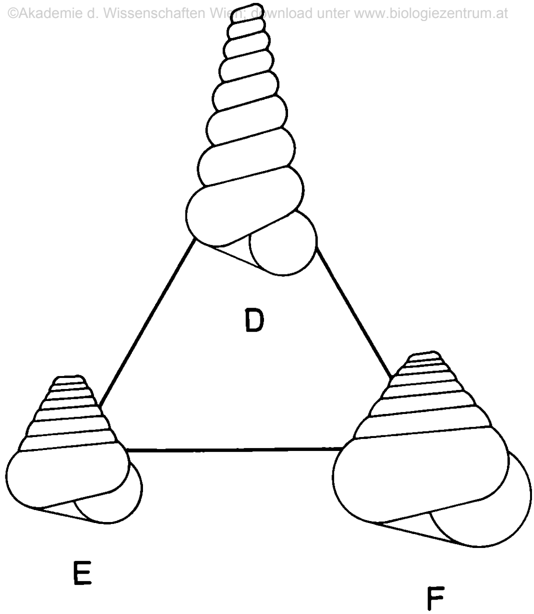
anzutreffen. Allein aus der Randverteilung des Merkmals Expansionsrate, die ungefähr einer Normalverteilung entspricht (vgl. Abb. 11 c), wäre diese Differenzierung nicht zu erkennen gewesen.

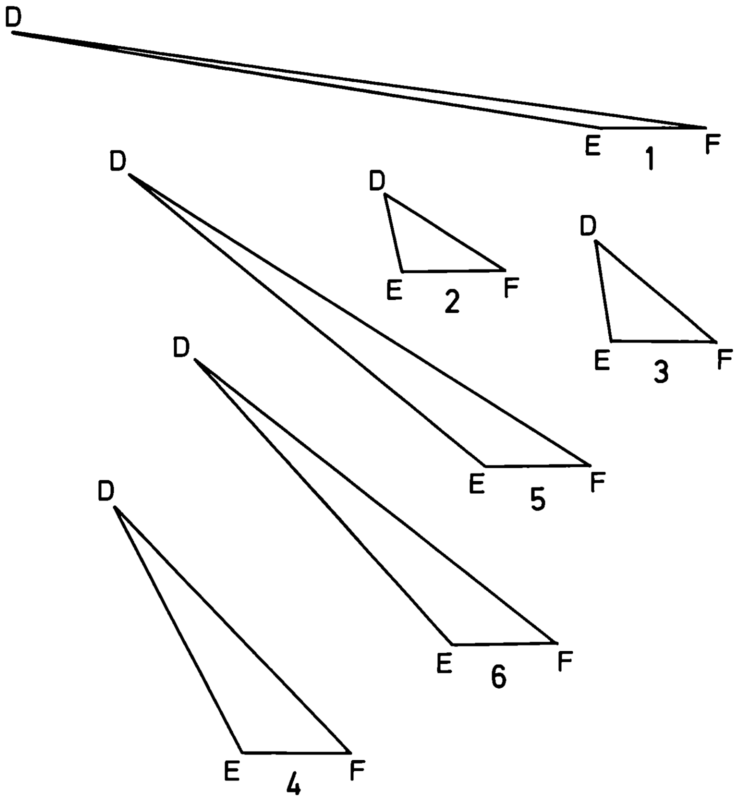
Durch die subjektive Angleichung der Wertebereiche von Expansions- und Versetzungsrate erfährt der Merkmalsraum eine starke Deformierung. Aber auch eine einfache Standardisierung erbringt keine besseren Resultate. Bei beiden Normierungsformen wird keine Rücksicht auf den Informationsgehalt genommen, der durch die Form der Randverteilungen vorgegeben ist. Im oben angeführten Beispiel ist dies besonders deutlich. Die Versetzungsrate zeigt eine deutliche Gliederung in Klassen. In der Randverteilung der Expansionsrate sind diese Hinweise auf Klassenstrukturen nicht zu erkennen (vgl. Abb. 11 c). Bei einer Standardisierung geht der Informationsvorsprung, der im Merkmal Versetzungsrate gegeben ist, vollständig verloren. Die Auswirkungen sollen in der Folge erläutert werden:

Die Dreiecke mit der Ziffer 1 der Abb. 12 und 13 repräsentieren die Ähnlichkeitsbeziehungen der Punkte A–F im ursprünglichen, von den Variablen W und T aufgespannten Merkmalsraum. Durch die ungleichen Wertebereiche kommt den Differenzen in der Expansionsrate fast keine Bedeutung zu, was sich in der extremen Schiefe der beiden Dreiecke ausdrückt. Mit der Ziffer 3 sind in beiden Abbildungen die Ähnlichkeitsrelationen dargestellt, die sich aus der einfachen Standardisierung des W-T-Merkmalsraumes ergeben. Besonders das Dreieck in Abb. 13 zeigt die starke Verzerrung der morphologischen Distanzen, die so weit geht, daß die völlig anders gestaltete Form D dem Individuum E plötzlich näher steht als dieses dem morphologisch sehr nahe verwandten Gehäuse F. Ähnliche Ergebnisse, vielleicht etwas prägnanter, bringt der Vergleich der Individuen A, B und C (Abb. 12/3), wo durch die Standardisierung das anders gestaltete Individuum C der Form B wesentlich näher steht als dieses dem Gehäuse A. Die Ähnlichkeitsbeziehungen stehen somit im krassen Widerspruch zur subjektiven Beurteilung von Ähnlichkeiten. Gleichzeitig beweisen sie, daß kritiklose Apriori-Standardisierungen extreme Verzerrungen bewirken können. Eine auf solchen Ähnlichkeitsbeziehungen basierende Klassifikation kann nur falsche Ergebnisse bringen.

Abb. 13: Vergleiche der Ähnlichkeitsrelationen zwischen den Individuen D, E und F in den unterschiedlich transformierten Merkmalsräumen, die von den Variablen Expansionsrate (W) und Versetzungsrate (T) aufgespannt werden.
Ähnlichkeitsrelationen als metrische Distanzen im
1. ursprünglichen Merkmalsraum (Abb. 11a),
2. transformierten Merkmalsraum (Abb. 11b),
3. einfach standardisierten Merkmalsraum,
4. klassenspezifisch standardisierten Merkmalsraum,
5. komponentenspezifisch standardisierten Merkmalsraum mit ungleicher Komponentenzahl in den Merkmalen,
6. komponentenspezifisch standardisierten Merkmalsraum mit gleicher Komponentenzahl in den Merkmalen.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at





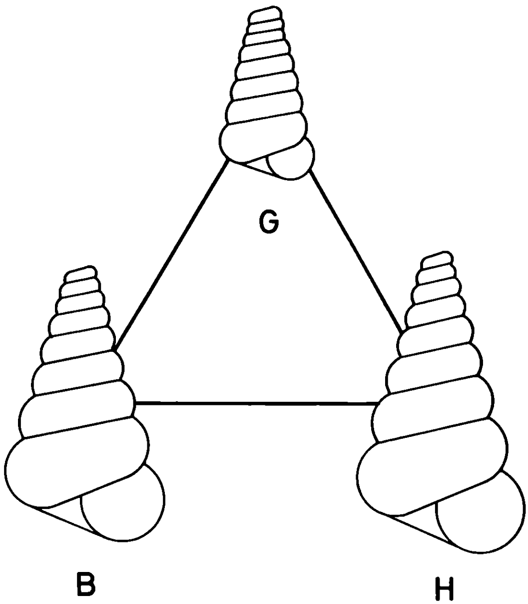
©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

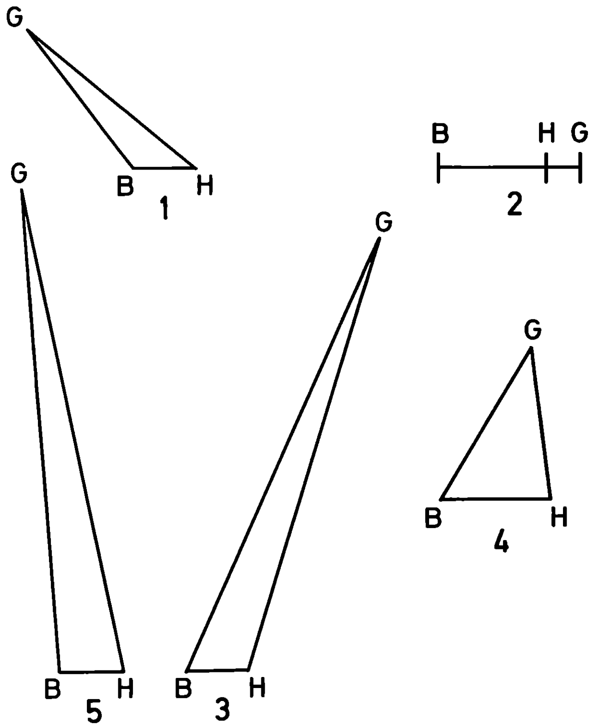
Sowohl in Abb. 12 als auch in Abb. 13 stellen die Dreiecke mit der Ziffer 4 die Konfiguration von Ähnlichkeitsrelationen dar, wie sie durch eine klassenspezifische Standardisierung hervorgerufen werden. Obwohl von der Population mit den niedrigen Gehäusen die extremen Formen (Individuen E und F) genommen wurden, sind sie, wie an den Distanzen im Dreieck 4 der Abb. 13 erkennbar, einander wesentlich ähnlicher als im Einzelvergleich dem Individuum D der Klasse mit den hohen Gehäusen. Diese Relationen stimmen mit dem subjektiven morphologischen Befund überein. Eine solche Kongruenz ist auch beim Vergleich der Morphologie der Gehäuse A, B und C mit den Ähnlichkeitsbeziehungen des Dreieckes Abb. 12/4 gegeben. Auch hier entsprechen die Distanzen zwischen den Punkten den Ähnlichkeitsrelationen, die man bei einer subjektiven Analyse erstellen würde. Daraus folgt, daß eine klassenspezifische Standardisierung der Merkmalswerte die beste Konfiguration der Individuen in einem Merkmalsraum bewirkt.

Eine klassenspezifische Standardisierung hat auch noch den Vorteil, daß sie, im Gegensatz zur Darstellung im ursprünglichen oder einfach standardisierten Merkmalsraum, die Verteilungsformen der Klassen berücksichtigt. Auf welche Weise sich dies vorteilhaft auswirken kann, sollen die folgenden Vergleiche zeigen: Nimmt man die Population mit den schlanken Individuen und vergleicht die mehr oder minder extremen Individuen G und H (vgl. Abb. 6) mit dem Individuum B, das der Form H stark ähnelt, ist im ursprünglichen Merkmalsraum, der von der Gehäusebreite und -höhe aufgespannt wird, die Ähnlichkeitsrelation durch das Dreieck Abb. 14/1 charakterisiert. In den Höhe/Breite-Quotienten wird jedoch eine völlig entgegengesetzte Ähnlichkeitsbeziehung geliefert (Abb. 14/2). Hier ist das Gehäuse B von der morphologisch fast identischen Form H wesentlich weiter entfernt als das morphologisch deutlich verschiedene Individuum G von H. Im Merkmalsraum, der von den unstandardisierten Variablen W und T aufgespannt wird, ergeben sich Relationen, wie sie im dritten Dreieck der Abb. 14 dargestellt werden. Diese Konfiguration steht im guten Einklang mit den subjektiven Ähnlichkeitsurteilen, nur würde der morphologisch versierte Betrachter sagen, daß bei einer intensiven Studie der drei Gehäuseformen das Individuum G der Form B etwas näher als der Form H stehen müßte.

Abb. 14: Vergleiche der Ähnlichkeitsrelationen zwischen den Individuen einer Klasse (B, G und H) in den unterschiedlich transformierten Merkmalsräumen.
Ähnlichkeitsrelationen als metrische Distanzen im

1. ursprünglichen Merkmalsraum, der von den Variablen Gehäusehöhe (y) und Gehäusebreite (x) aufgespannt wird (vgl. Abb. 6),
2. eindimensionalen Merkmalsraum des Quotienten y/x (vgl. Abb. 9a),
3. ursprünglichen Merkmalsraum, der von den Variablen W (= Expansionsrate) und T (= Versetzungsrate) aufgespannt wird (vgl. Abb. 11a),
4. einfach standardisierten Merkmalsraum, der von den Variablen W und T aufgespannt wird,
5. klassenspezifisch standardisierten Merkmalsraum, der von den Variablen W und T aufgespannt wird.


G
B
H


G
B
H
1
B
H
G
2
G
B
H
5
G
B
H
3
G
B
H
4

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Diese Beobachtungen stimmen völlig mit den Ähnlichkeitsrelationen überein, die eine klassenspezifische Standardisierung bringt (vgl. Abb. 14/5).

Im ursprünglichen Merkmalsraum der Höhen- und Breitenmaße stehen die Individuen B und G in den Extrembereichen der Verteilungsellipse mit den schlanken Gehäusen (vgl. Abb. 6). Die morphologische Distanz ist sehr groß, wie man aus dem langen Schenkel des ersten Dreieckes der Abb. 15 zwischen den beiden Punkten erkennen kann. Diese Distanz ist größer als die Verbindungen der Punkte B und G zum mittleren Punkt der Population mit den gedrungenen Gehäusen (Individuum C). Diese Ähnlichkeitsrelationen stimmen jedoch nicht mit denen der subjektiven Beurteilung von Ähnlichkeiten überein. Ein Quotienten-Vergleich liefert wesentlich bessere Resultate (Abb. 15/2). Ähnliche Ergebnisse müßte auch ein Vergleich mit den Merkmalen Expansions- und Versetzungsrate liefern (Abb. 15/3), insbesondere da der Parameter T eine ähnliche Diskriminanz der Klassen wie der Quotient liefert. Wegen der Unterbewertung der Differenzen, die in der Expansionsrate liegen, muß eine Standardisierung durchgeführt werden. In einfacher Form (Abb. 15/4) bringt sie wieder eine totale Deformation der Ähnlichkeitsbeziehungen, hingegen ergeben sich bei einer klassenspezifischen Standardisierung Konfigurationen, die mit dem subjektiven Empfinden von Ähnlichkeiten absolut konform gehen (Abb. 15/5).

Anhand der oben aufgezeigten Beispiele wurde vorgeführt, daß nur eine klassenspezifische Standardisierung von Merkmalswerten, die allerdings die Morphologie von Individuen erschöpfend erklären müssen, befriedigende und mit den subjektiven Urteilen über Ähnlichkeitsbeziehungen übereinstimmende Ergebnisse liefert.

Methodik

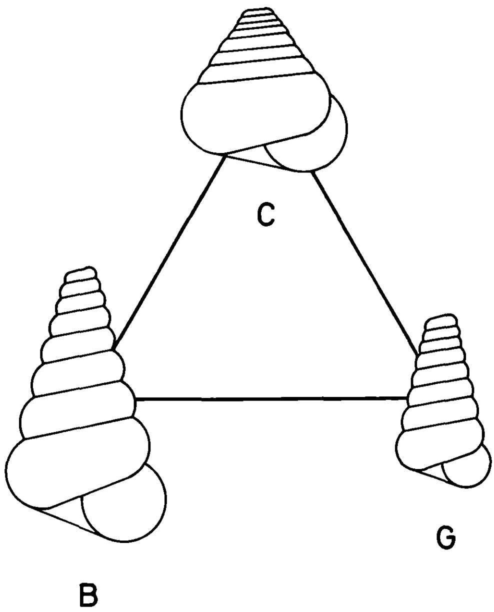
In diesem Kapitel wird erläutert, auf welche Weise eine an die klassenspezifische Standardisierung angenäherte Variante der Normierung durchzuführen ist. Am Beginn einer automatischen Klassifikation liegen die Merkmale mit ihren Häufigkeitsverteilungen vor. Sie repräsentieren die Randverteilungen einer multivariaten Häufigkeitsverteilung. Die Formen der Randverteilungen liefern bereits Hinweise auf die Struktur der Objektmenge im Merkmalsraum. Ist die Information

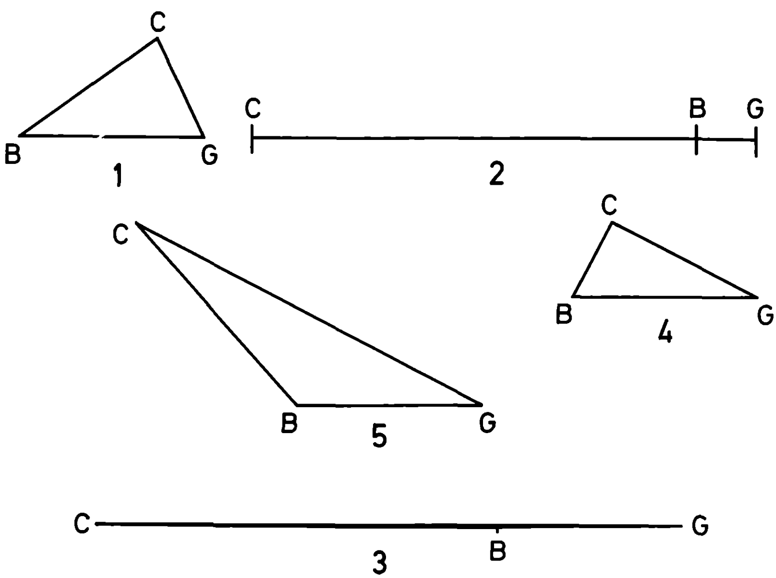
---

Abb. 15: Vergleiche der Ähnlichkeitsrelationen zwischen den extremen Individuen einer Klasse (B und G) mit einem mittleren Exemplar der anderen Klasse (Individuum C). Ähnlichkeitsrelationen als metrische Distanzen im

1. ursprünglichen Merkmalsraum, der von den Variablen Gehäusehöhe (y) und Gehäusebreite (x) aufgespannt wird (vgl. Abb. 6),
2. eindimensionalen Merkmalsraum des Quotienten y/x (vgl. Abb. 9a),
3. ursprünglichen Merkmalsraum, der von den Variablen W (= Expansionsrate) und T (= Versetzungsrate) aufgespannt wird (vgl. Abb. 11a),
4. einfach standardisierten Merkmalsraum, der von den Variablen W und T aufgespannt wird,
5. klassenspezifisch standardisierten Merkmalsraum, der von den Variablen W und T aufgespannt wird.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at





©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

über die Struktur vollständig in den Randverteilungen enthalten, wobei diese deutlich getrennte Häufigkeitskonzentrationen aufweisen (z. B. Abb. 11 c, Versetzungsrate T), lassen sich die Klassen bereits in den Randverteilungen erkennen, und eine weitere Klassifikation erübrigt sich. Auf diesen Überlegungen basiert die Vorgangsweise einer monothetischen Klassifikation.

In den meisten Fällen sind diese Randverteilungen jedoch mehrgipfelige Mischverteilungen von Klassen verschiedenster Größe, die unter ungünstigen Voraussetzungen sogar einheitliche, unimodale und oft auch Normalverteilungen sein können. Anhand solcher Mischverteilungen ist die Klassenstruktur der Objektmenge aus den Formen der Randverteilungen nicht eindeutig zu erkennen. Ein einzelnes Individuum läßt sich keiner bestimmten Klasse zuordnen, wie es bei einer monothetischen Klassifikation zu fordern wäre.

Übereinstimmende Formen von Randverteilungen können durch unterschiedliche Konfigurationen der Objektmenge im Merkmalsraum hervorgerufen werden. In Abb. 16 wird zur Illustration wieder ein zweidimensionaler Raum verwendet. Die Form der beiden Randverteilungen ist vorgegeben, wobei die Variable x eine Normalverteilung, die Variable y eine schiefe Verteilung zeigt. Aus der Vielzahl von Möglichkeiten der Konfiguration des Merkmalsraumes, wie er durch die beiden Randverteilungen charakterisiert wird, sollen drei Varianten vorgestellt werden. In den ersten beiden Fällen sind die Verteilungen einheitlich und symmetrisch, wobei die Symmetrieachse der ersten Variante normal zur x-Achse gelegen ist, im zweiten Beispiel jedoch eine schräge Lage im Merkmalsraum aufweist. Der dritte Fall zeigt, daß dieselben Randverteilungen auch durch zwei homogene Teilmengen hervorgerufen werden können. Dies möge als Beweis dienen, daß die Randverteilungen nur einen bestimmten Teil der Information über die

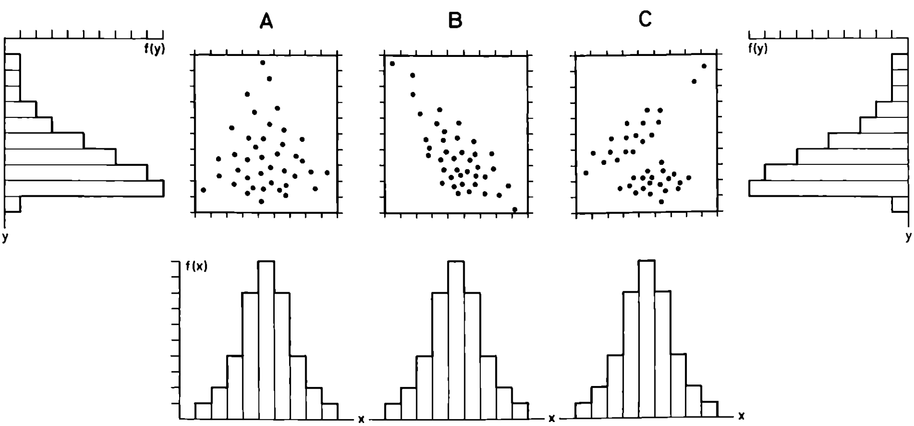


Abb. 16: Unterschiedliche Möglichkeiten der Punktekonfiguration in einem zweidimensionalen Merkmalsraum bei konstanten Randverteilungen.

Struktur einer Objektmenge im mehrdimensionalen Merkmalsraum beinhalten können.

Trotzdem beschränken die Randverteilungen durch ihre Verteilungsformen die Möglichkeiten der Objektkonfiguration im Merkmalsraum. Auf diesen Teil der Information, deren Prozentanteil an der Gesamtinformation der Struktur einer Objektmenge nur schwer abzuschätzen ist, sollte bei der Beschreibung des Merkmalsraumes näher eingegangen werden.

Aus den in den vorigen Kapiteln genannten Gründen muß eine Normierung der Merkmalswerte unbedingt gefordert werden, wobei eine Standardisierung allen anderen Normierungsverfahren vorzuziehen ist. Eine einfache Standardisierung läßt sich jedoch sinnvoll nur an normalverteilten Variablen durchführen. Liegen keine Normalverteilungen vor, können die Häufigkeitsverteilungen durch geeignete Transformationen in Normalverteilungen überführt werden (vgl. Box & Cox, 1964; Sokal & Rohlf, 1969, 380 ff.). Ergeben auch diese Transformationen keine Normalverteilungen, wäre folgender Weg einzuschlagen:

Jede Häufigkeitsverteilung kann als mathematische Funktion f(x) beschrieben werden, man spricht von einer Häufigkeitsfunktion (vgl. Kreyszig, 1968, 24). Ist nun eine Häufigkeitsfunktion g(x) gegeben, so versucht man eine Schar von Häufigkeitsfunktionen $f(x,\alpha,\beta)$ von einer gewissen analytischen Form zu finden, so daß sich g(x) als Überlagerung von Funktionen $f(x,\alpha_l,\beta_l)$ schreiben läßt (vgl. Medgyessy, 1977, 17 ff.):

$$g(x) = \sum_{l=1}^{h} p_l f(x, \alpha_l, \beta_l) \quad (\alpha_l \varepsilon A, \beta_l \varepsilon B)$$

In dieser Form stellt $f(x, \alpha,\beta)$ eine Funktion dar, die durch die Parameter α und β charakterisiert ist. Mit h ist die Zahl der Komponenten, mit p_l das Gewicht der l-ten Komponente indiziert.

Dies gilt für den allgemeinen Fall der Aufgliederung von Funktionen in Komponenten. Beim vorliegenden Problem der Standardisierung nicht-normalverteilter stetiger Häufigkeitsfunktionen ist die analytische Form der Komponenten determiniert. Zerlegt man die Häufigkeitsfunktion in normalverteilte Komponenten (vgl. Medgyessy, 1977, 103), so gewinnt sie folgendes Aussehen:

$$g(x) = \sum_{l=1}^{h} p_l \frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma_l}\right)^2\right]}{\sigma_l\sqrt{2\pi}} \quad (-\infty < x < \infty, \sigma_l > 0) \qquad (6)$$

Die Parameter μ_l und σ_l sind die Mittelwerte und Standardabweichungen der Komponenten. Mit ihnen läßt sich eine Standardisierung der Merkmalswerte anhand der Formel 4 durchführen, die „komponentenspezifische Standardisierung" genannt werden soll.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Die Zerlegung von mehrgipfeligen oder asymmetrischen Häufigkeitsverteilungen in normalverteilte Komponenten bedeutet nicht, daß diese mit den Klassen der Objektmenge im Merkmalsraum bereits identisch sind (es wäre dann eine klassenspezifische Standardisierung, für die aber die Klassen bereits bekannt sein müssen). Die Komponenten dienen nur dazu, die nicht-normalverteilten Häufigkeiten zu charakterisieren und mit den Parametern der einzelnen normalverteilen Komponenten die Merkmalswerte zu standardisieren.

Eine Zerlegung von Häufigkeitsverteilungen in normalverteilte Komponenten läßt sich mit unterschiedlichen Verfahren durchführen. Neben graphischen Methoden (z. B. BHATTACHARYA, 1967) gibt es komplexere analytische Verfahren, die einen hohen Rechenaufwand erfordern und nur mehr mit Hilfe elektronischer Rechenanlagen durchzuführen sind (vgl. GREGOR, 1969; MEDGYESSY, 1977).

Eine Standardisierung mit normalverteilten Komponenten führt zu einer unterschiedlichen Gewichtung der Merkmale im standardisierten Merkmalsraum. In einer einfachen Standardisierung werden die Wertebereiche auf annähernd gleiche Größe normiert (Grenzen etwa zwischen -2 und $+2$). Bei der komponentenspezifischen Standardisierung bestimmen die Individuen, die außerhalb des Streuungsbereiches einer Komponente zu liegen kommen, die Größe des Wertebereichs eines Merkmals (vgl. klassenspezifische Standardisierung). Als Maßeinheit dafür gilt die Standardabweichung der Komponente, mit der die Standardisierung durchgeführt wird. Neben dieser Ausdehnung der Wertebereiche, die durch die Lage der anderen Komponenten bedingt ist, erfolgt noch eine zusätzliche Vergrößerung durch die Zahl der Komponenten, mit denen jeweils eine eigene Normierung durchzuführen ist. Beispielsweise führt eine Zerlegung eines Merkmals in drei Komponenten zu einer Erhöhung der Dimensionalität des Merkmalsraumes um den Faktor 3. Obwohl nur ein einziges Merkmal zerlegt wurde, unterscheiden sich die Wertebereiche der drei neuen Merkmalsdimensionen. Diese Unterschiede werden durch die Differenzen in den Verteilungsparametern der einzelnen Komponenten bewirkt. Liegen die Komponenten eng nebeneinander und überschneiden sich die Streuungsbereiche sehr stark, so führt dies zu keiner besonderen Erweiterung des Wertebereiches. Ist jedoch eine Komponente gegenüber der anderen durch einen geringen Streuungsbereich und eine extreme Lage ausgezeichnet, so bewirkt eine Standardisierung mit ihren Parametern eine beträchtliche Erweiterung des Merkmalsraumes, was gleichzeitig eine Gewichtung des aufgeteilten Merkmals bedeutet. Eine Erweiterung des Wertebereiches ist mit einer Vergrößerung der Distanzen verbunden; Unterschiede in diesem Merkmal bekommen mehr Gewicht.

Die Anordnung der normalverteilten Komponenten eines Merkmals, ihre Lage zueinander und die Größe der Streuungsbereiche sind daher für die Gewichtung des Merkmals von großer Bedeutung. Da die Komponenten aus der Häufigkeitsverteilung eines Merkmals gewonnen

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

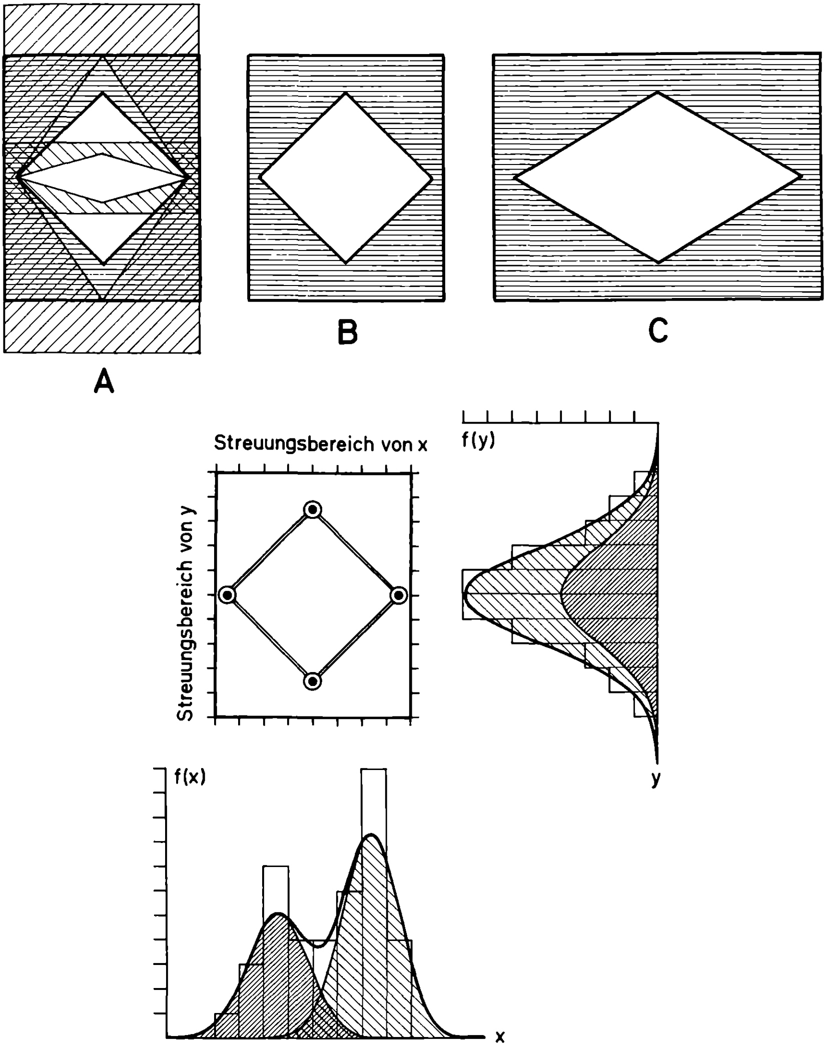


Abb. 17: Aufgliederung von Randverteilungen in normalverteilte Komponenten und Gewichtung der Merkmale durch die Erweiterung des Merkmalsbereiches bei einer komponentenspezifischen Standardisierung.
Die empirische Randverteilung (Histogramm) des Merkmals x läßt sich in zwei normalverteilte Komponenten auflösen (Funktionskurven mit unterschiedlich schraffierten Flächen). Infolgedessen muß auch das normalverteilte Merkmal y in zwei Komponenten (mit identischen Verteilungsparametern) gegliedert werden.
Durch unterschiedliche Maßeinheiten hervorgerufene Differenzen in den Wertebereichen der beiden Merkmale (Fig. A) werden mit einer Standardisierung ausgeglichen. Die einfache Standardisierung bringt eine annähernd gleiche Dimensionierung der Wertebereiche (Fig. B), während die komponentenspezifische Standardisierung das deutlicher strukturierte Merkmal (Variable x) durch eine Erweiterung des Wertebereiches gewichtet (Fig. C).

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

werden und diese in ihrer Form vollständig zu charakterisieren vermögen, ergibt sich das erwünschte Resultat, daß bei einer komponentenspezifischen Standardisierung die Form der Häufigkeitsverteilung in die Klassifikation einbezogen wird.

Kann eine Variable mit einer mehrgipfeligen Häufigkeitsverteilung in eine bestimmte Zahl von normalverteilten Komponenten zerlegt werden, so wird der Merkmalsraum durch diese Komponenten erhöht. Dadurch gewinnt jedoch das Merkmal gegenüber den normalverteilten oder eingipfeligen Variablen ein allzu großes Gewicht. Es muß aus diesem Grund gefordert werden, daß alle Merkmale in die gleiche Zahl von Komponenten zerlegt werden. Bei einem normalverteilten Merkmal schafft diese Forderung keine Probleme, wie Abb. 17 verdeutlichen soll. In diesem Beispiel eines zweidimensionalen Merkmalsraumes läßt sich die erste Variable mit einer zweigipfeligen Häufigkeitsverteilung in zwei normalverteilte Komponenten zerlegen. Das zweite Merkmal (y) ist durch eine Normalverteilung ausgezeichnet. Diese muß, nach der obigen Forderung der gleichen Komponentenzahl in allen Merkmalen, gleichfalls in zwei Komponenten zerlegt werden. Man nimmt an, daß diese Häufigkeitsverteilung eine Summe zweier Normalverteilungen mit den selben Verteilungsparametern (gleiche Mittelwerte und gleiche Varianzen) ist. Kann ein Merkmal in eine bestimmte Zahl (h) von Komponenten zerlegt werden, dann läßt sich nach der obigen Überlegung ein normalverteiltes Merkmal gleichfalls in h-Komponenten aufgliedern, die alle die gleichen Verteilungsparameter besitzen.

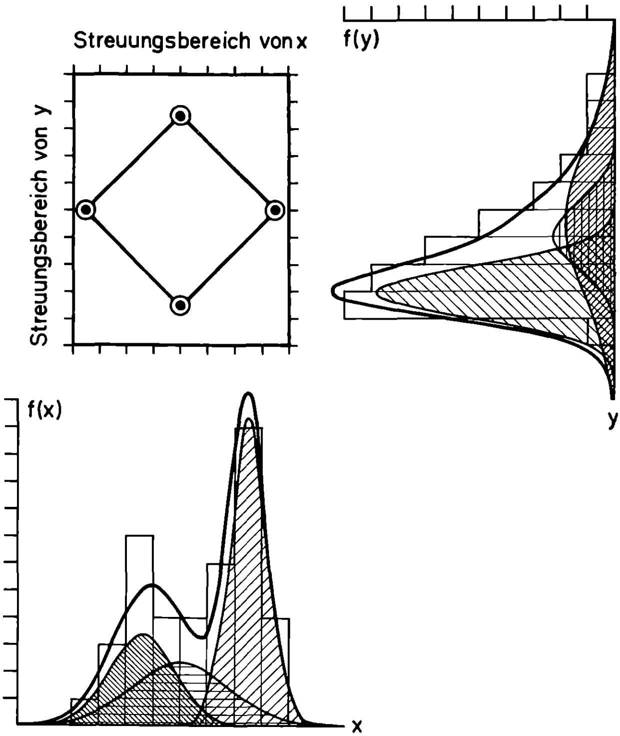
Ein Vergleich der Wertebereiche des Merkmalsraumes der Abb. 17, wie er sich nach einer komponentenspezifischen Standardisierung präsentiert, zeigt deutlich die erhöhte Gewichtung des Merkmals mit der deutlich strukturierten Randverteilung (großer Wertebereich) gegenüber der mehr oder minder unstrukturierten zweiten Variablen (geringer Wertebereich).

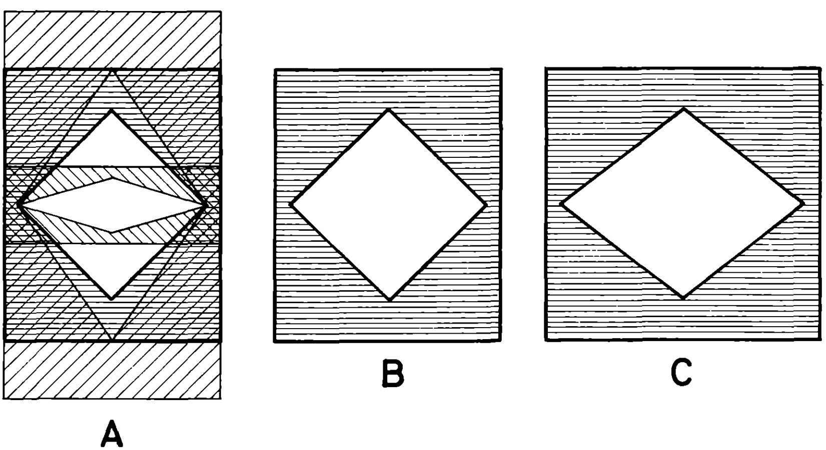
Daß die Zerlegung der normalverteilten Variablen in eine von den anderen Merkmalen vorgegebene Zahl durchgeführt werden muß, sollen die Beispiele 5 und 6 der Abb. 12 und 13 zeigen. Hier wurde im Beispiel 4 die klassenspezifische Standardisierung mit den Parametern der bereits vorgegebenen Teilmengen (= Klassen) der Objektmenge durchgeführt. Bei einer Klassifikation ist jedoch die Klassenstruktur von vornherein unbekannt, man kennt nur die Randverteilungen (vgl. Abb. 11 c), anhand derer eine

---

Abb. 18: Aufgliederung von Randverteilungen in normalverteilte Komponenten und Gewichtung der Merkmale durch die Erweiterung des Merkmalsbereiches bei einer komponentenspezifischen Standardisierung.
Die empirische Randverteilung (Histogramm) des Merkmals x ist identisch mit der Abb. 17, die Randverteilung der Variablen y zeigt jedoch eine schiefe Verteilung. Eine Aufgliederung dieses Merkmals erbringt drei normalverteilte Komponenten. Infolgedessen muß auch das Merkmal x mit der deutlichen Zweigipfeligkeit in drei Komponenten aufgelöst werden. Wie in Abb. 17 werden die durch unterschiedliche Maßeinheiten hervorgerufenen Differenzen in den Wertebereichen der Merkmale (Fig. A) mit einer Standardisierung ausgeglichen. Während die einfache Standardisierung eine annähernd gleiche Dimensionierung der Wertebereiche bewirkt (Fig. B), wird bei der komponentenspezifischen Standardisierung trotz gleicher Komponentenzahl das deutlicher strukturierte Merkmal (Variable x) durch einen größeren Wertebereich stärker gewichtet.

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at





©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

komponentenspezifische Standardisierung durchzuführen ist. In unserem Beispiel zeigt die Versetzungsrate T die deutliche Trennung der beiden Klassen, die Expansionsrate W ist jedoch durch eine eingipfelige Verteilung ausgezeichnet. Sie entspricht ungefähr einer Normalverteilung; die beiden Klassen lassen sich aus dem Histogramm nicht erkennen. Würde man bei einer komponentenspezifischen Standardisierung diese Variable nicht in zwei Komponenten zerlegen, so ergäben sich Ähnlichkeitsrelationen, wie sie im Beispiel 5 der Abbildungen 12 und 13 dargestellt sind. Teilt man die Häufigkeitsverteilung der Expansionsrate in zwei Komponenten mit identischen Verteilungsparametern und führt dann die Standardisierung durch, so erhöht sich der Merkmalsraum auf vier Dimensionen, und die Ähnlichkeitsbeziehungen (Abb. 12/6 und Abb. 13/6) nähern sich sehr stark den Relationen, die durch die klassenspezifische Standardisierung gewonnen wurden (Abb. 12/4 und Abb. 13/4).

Da die Merkmale jeweils in die selbe Zahl von Komponenten zu zerlegen sind, bestimmt das Merkmal mit der höchsten Komponentenzahl den Grad der Zerlegung. Die Schätzverfahren für eine Zerlegung von Häufigkeitsverteilungen in eine vorgegebene Zahl von Komponenten basieren auf der Maximum-Likelihood-Methode (vgl. z. B. FANGMEYER, 1964).

Auch wenn alle Variablen die selbe Komponentenzahl aufweisen, bewirken die Differenzen in der Lage und Streuung der Komponenten eine unterschiedliche Gewichtung der Merkmale. Im Beispiel Abb. 18 ist bei der ersten Variablen die gleiche zweigipfelige Häufigkeitsverteilung wie in der Abb. 17 zu sehen. Die zweite Variable ist jedoch durch eine schiefe Verteilung gekennzeichnet, die sich in drei normalverteilte Komponenten zerlegen läßt. Somit muß auch das Merkmal mit der deutlichen Zweigipfeligkeit in drei Komponenten zerlegt werden. Obwohl die beiden Variablen in die gleiche Zahl von Komponenten zerlegt wurden, bewirkt die unterschiedliche Form der Häufigkeitsverteilung eine verschiedene Gewichtung der Merkmale. Dies drückt sich in den Dimensionen der Wertebereiche aus. Die Variable mit der deutlichen zweigipfeligen Verteilung gewinnt durch die komponentenspezifische Standardisierung einen größeren Wertebereich als das Merkmal mit der schiefen Häufigkeitsverteilung (vgl. Abb. 18 C). Dies bedeutet, daß Variablen mit einer deutlicheren Struktur bei einer komponentenspezifischen Standardisierung auch stärker gewichtet werden. Eine einfache Standardisierung würde die Wertebereiche der beiden Merkmale egalisieren (vgl. Abb. 18 B).

In diesem Artikel wurde bislang nur die Normierung von quantitativen Merkmalen besprochen. Werden auch qualitative Merkmale in eine Klassifikation einbezogen, muß auf alle Fälle eine Normierung der Merkmale in irgendeiner Form durchgeführt werden. Bei einer Vernachlässigung der Normierung ist der Wertebereich eines qualitativen Merkmals durch das Intervall [0,1] gegeben und kann nicht direkt mit den Wertebereichen der unnormierten quantitativen Variablen verglichen werden, die extrem schwanken können.

Wird bei quantitativen Variablen eine Normierung auf den Merkmalsbereich durchgeführt, erübrigt sich diese bei den qualitativen, dichotomen Variablen, da deren Merkmalsausprägungen eben die

Grenzen des Intervalles [0,1] sind (zur Verwendung der Normierung auf den Merkmalsbereich bei gemischten Daten vgl. GOWER, 1971).

Die zweite Form der Normierung ist die Standardisierung. Sie läßt sich auch bei binären (= dichotomen) Variablen durchführen (vgl. SNEATH & SOKAL, 1973, 165). Es wird dabei jedoch keine Normalverteilung in eine standardisierte Normalverteilung übergeführt, sondern eine Zweipunktverteilung in eine Verteilung mit dem Mittelwert 0 und der Standardabweichung 1 transformiert. Die Zweipunktverteilung ist eine Sonderform der Binomialverteilung der Form:

$$f(x) = \binom{n}{p} p^x q^{n-x}$$

$$\text{mit } n = 1,\ \mu = p,\ \sigma = \sqrt{pq} \text{ und } p + q = 1.$$

Da die Merkmalswerte bei einer Zweipunktverteilung nur die Endpunkte des Intervalls [0,1] einnehmen können, bestimmt die Distanz zwischen den Intervallgrenzen den Grad der Ähnlichkeit. Bei unnormierten Daten bedeutet eine Distanz von 0 denselben Merkmalswert, eine Distanz von 1 die alternativen Merkmalsausprägungen. Führt man eine Standardisierung durch

$$x^*_{ik} = \frac{x_{ik} - p_k}{\sqrt{p_k q_k}}, \tag{7}$$

dann hängt die Distanz zwischen den beiden alternativen Merkmalsausprägungen vom Mittelwert der Verteilung, das heißt vom Prozentanteil der beiden Gruppen mit den unterschiedlichen Merkmalswerten, ab. Nur im Fall einer Gleichverteilung, wenn also beide Ausprägungen mit 50 % der Gesamtzahl vertreten sind, ergibt sich für das Intervall die Größe 2. Alle Abweichungen von diesen Prozentanteilen vergrößern die Abstände der Intervallgrenzen. Beispielsweise gehen besonders kleine Gruppen mit einem Prozentanteil $p < 0{,}1$ mit größerem Gewicht in eine Klassifikation ein. Bei einem Wert von $p = 0{,}1$ beträgt die Distanz zwischen den beiden Alternativen bereits 6 und vergrößert sich bei noch kleineren Prozentanteilen beträchtlich.

Wird eine einfache Standardisierung an qualitativen Variablen durchgeführt, erreicht man den Effekt, der bei quantitativen Variablen erst durch eine klassenspezifische Standardisierung mit einer zusätzlichen Gewichtung durch die Klassenanteile erreicht wird. Ähnlich der einfachen Standardisierung qualitativer Variablen gewichtet die klassen- oder komponentenspezifische Standardisierung gleichfalls Komponenten mit geringer Varianz und extremer Lage auf der Merkmalsskala.

Gehen in eine Ähnlichkeitsanalyse auch qualitative-dichotome Variable gemeinsam mit quantitativen Merkmalen ein und wird eine komponentenspezifische Standardisierung durchgeführt, so bestimmt in

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

diesem Fall die Variable mit der höchsten Komponentenzahl den Grad der Aufteilung. Die vorliegenden Zweipunktverteilungen müssen dann als Summe von identischen Zweipunktverteilungen gedeutet werden, deren Zahl vom Grad der Gliederbarkeit der quantitativen Merkmale bestimmt wird. Wenn eine Gliederung in h-Komponenten möglich ist, müssen bei einem qualitativen Merkmal auch h-Zweipunktverteilungen zur Standardisierung herangezogen werden.

Abschließend sei bemerkt, daß eine komponentenspezifische Standardisierung durch ihre Bevorzugung von deutlich strukturierten Häufigkeitsverteilungen Merkmale hervorhebt und besonders gewichtet, die auch bei einer intuitiven Ähnlichkeitsanalyse durch ihre deutliche Struktur bevorzugt werden.

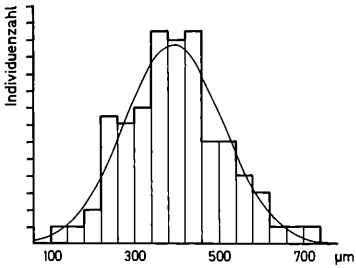
Beispiel

Die Effizienz der komponentenspezifischen Standardisierung soll an einem Beispiel verdeutlicht werden. Als Untersuchungsobjekte dienten 86 Gehäuse von Einzellern (Foraminiferen) der Gattung *Ichthyolaria*. Die vier untersuchten Arten *Ichthyolaria sulcata* (BORNEMANN), *I. terquemi* (D'ORBIGNY), *I. squamosa* (TERQUEM & BERTHELIN) und *I. densicostata* (HOHENEGGER) stammen aus dem Unteren Jura Mitteleuropas. Sie sind in ihrem Aussehen mehr oder minder deutlich verschieden. Die Artengruppe wurde von HOHENEGGER 1980 und 1981 eingehend untersucht, wobei zahlreiche Messungen an den Gehäusen durchgeführt wurden. Aus der Zahl von 20 Parametern, die die Gehäuseform geometrisch fast vollständig zu charakterisieren vermögen, wurden zur Verdeutlichung des Verfahrens 6 Variable ausgewählt. Es muß betont werden, daß es sich bei den Variablen um wachstumsunabhängige Merkmale handelt. Außerdem sind sie durch ihre hohen diskriminatorischen Eigenschaften zwischen den Gruppen ausgezeichnet (vgl. HOHENEGGER, 1981, 24 ff.). Folgende Merkmale gingen in die Analyse ein:

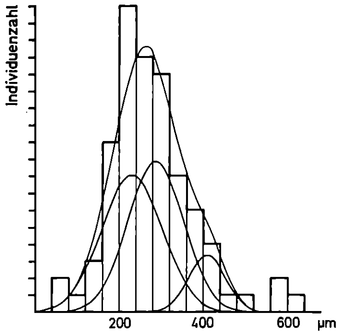
1. Der Durchmesser der Anfangskammer.
2. Der Wendepunkt in der Anstiegsrate des Gehäusewachstums.
3. Das prozentuelle Verhältnis der Kammerdicke zur Kammerbreite an der Stelle des größten Gehäusewachstums.
4. Der Winkel, den die Kammern mit ihren Wänden bei den Kammeröffnungen und der Mündung einschließen.
5. Die Distanz zwischen den Rippen.
6. Die Präsenz bzw. das ,,Unterdrücken" von lateralen Rippen.

Abb. 19: Empirische und theoretische Häufigkeitsverteilungen von sechs Merkmalen der liassischen Foraminiferengattung *Ichthyolaria* (86 Individuen).
Unter den fünf quantitativen Variablen zeigt allein das Merkmal Rippendistanz eine statistisch signifikante Abweichung von einer Normalverteilung. Seine Verteilungsform läßt sich als Summe dreier normalverteilter Komponenten erklären. Die Häufigkeitsverteilung des Merkmals Rippenunterdrückung, einer qualitativen Variablen, ist als Zweipunktverteilung dargestellt.

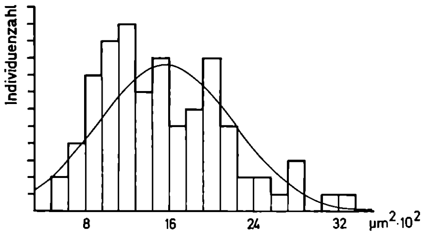
PROLOCULUSGRÖSSE



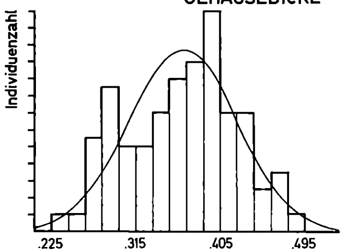
RIPPENDISTANZ



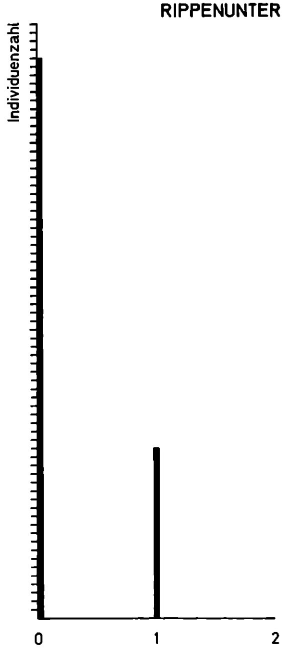
WACHSTUMSWENDEPUNKT



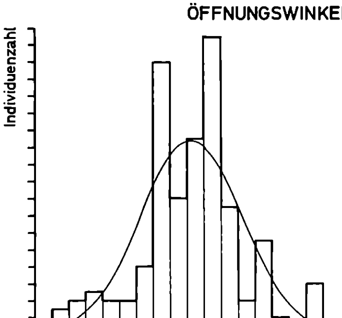
GEHÄUSEDICKE



RIPPENUNTERDRÜCKUNG



ÖFFNUNGSWINKEL



Bei den Merkmalen 1 bis 5 handelt es sich um quantitative Variable, das 6. Merkmal, eine qualitative Variable, läßt sich in dichotomer Form darstellen.

In der Folge wurden die Variablen 1–5 auf Normalverteilung überprüft, wobei als statistisches Verfahren der Kolmogoroff-Smirnov-Test herangezogen wurde. Als Signifikanzschwelle wurden 5 % Irrtumswahrscheinlichkeit angenommen. Mit Ausnahme der Rippendistanz waren bei allen Merkmalen die Abweichungen von einer Normalverteilung nicht signifikant.

Wie aus Abb. 19 zu ersehen ist, konnte die empirische Häufigkeitsfunktion des Merkmals Rippendistanz in drei normalverteilte Komponenten zerlegt werden. Als Methode zur Zerlegung wurde das graphische Verfahren von BHATTACHARYA, 1967, herangezogen.

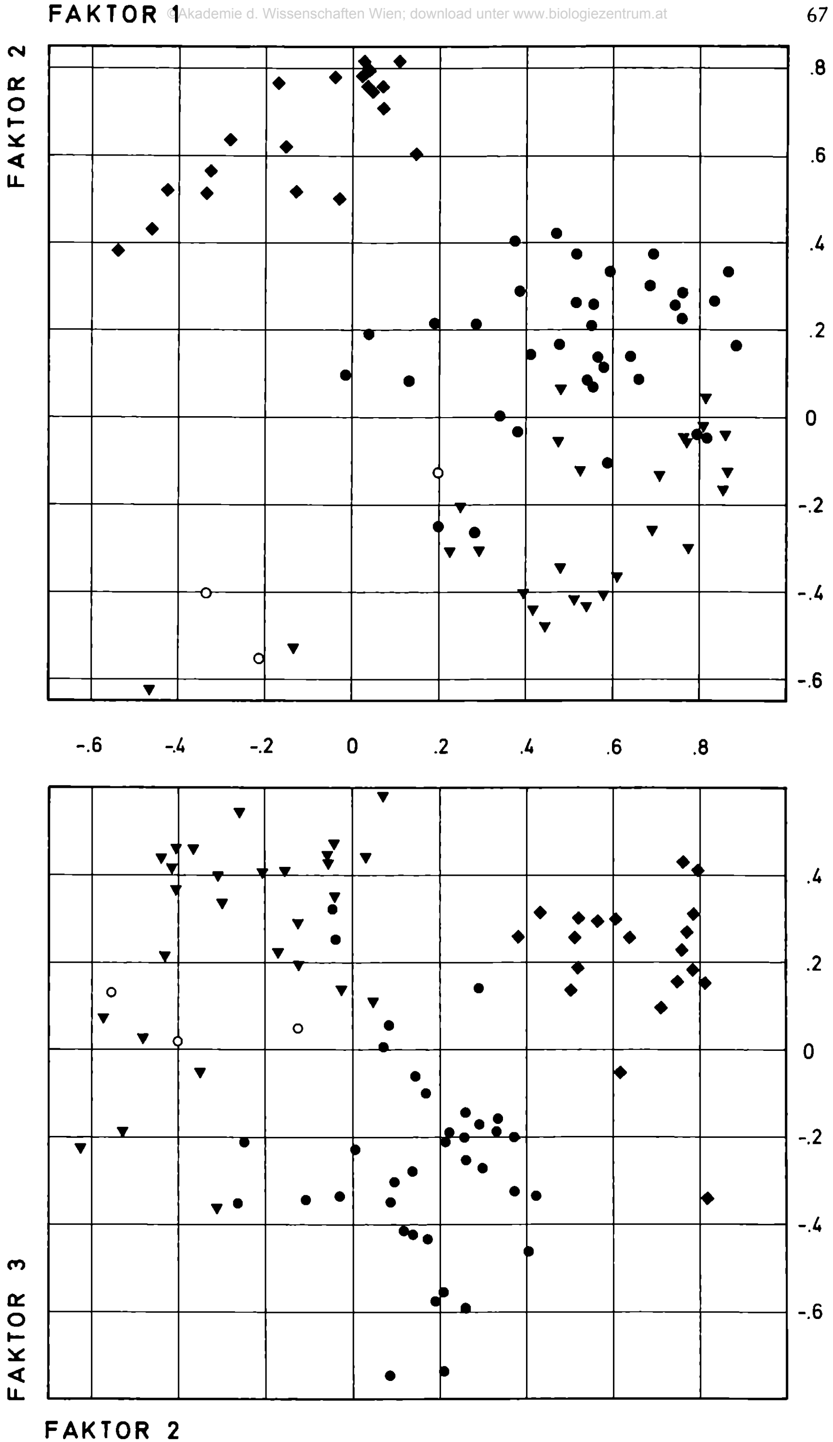
Die Lage und Varianz der Komponenten des Merkmals Rippendistanz variieren nicht sehr stark, so daß keine extreme Gewichtung der Merkmalswerte zu erwarten ist. Da die anderen Merkmale normalverteilt sind und nur ein Merkmal eine eher geringfügige Abweichung von einer Normalverteilung zeigt, ist zu erwarten, daß die Klassifikationen mit einfach und komponentenspezifisch standardisierten Merkmalswerten nicht zu stark differieren.

Das Resultat einer hierarchischen Klassifikation (UPGMA, vgl. SNEATH & SOKAL, 1973, 230 ff.) brachte bei den einfach standardisierten Merkmalen eine Zahl von neun Fehlklassifikationen. Hier soll unter dem Begriff „Fehlklassifikation" eine Zuordnung von Individuen zu einer Gruppe verstanden werden, die dem Empfinden eines Morphologen, d. h. der Intuition, völlig widerspricht. Mit einer komponentenspezifischen Standardisierung reduziert sich diese Zahl auf sechs Fehlklassifikationen. Zu erwähnen sei noch, daß die Zahl von falsch klassifizierten Individuen bei unstandardisierten Daten unüberschaubar wurde, es konnte keine „sinnvolle" Gruppierung erreicht werden.

Wie schon im Anfangskapitel erwähnt, sind für eine Klassifikation von Individuen die hierarchischen Verfahren ungeeignet. Darum wurde hier im Anschluß an diese Klassifikationsmethoden ein Verfahren eingesetzt, das in die Gruppe der multidimensionalen Skalierungen fällt. Es handelt sich dabei um eine Q-Mode-Faktorenanalyse (vgl. CATTELL, 1965). Da in eine Faktorenanalyse als Ähnlichkeitsmaße nur Korrelationskoeffizienten eingehen können, mußten die Distanzen in Ähnlichkeitsmaße transformiert werden, die dem Produktmomentkorrelationskoeffizienten äquivalent sind. Es wurde hier eine Transformation nach CATTELL, 1949, durchgeführt.

Abb. 20: Konfiguration von 86 Individuen der Gattung *Ichthyolaria* mit einfach standardisierten Merkmalswerten im dreidimensionalen Faktorenraum (Q-Mode-Analyse).
Ichthyolaria terquemi = schwarze Quadrate
Ichthyolaria sulcata = schwarze Kreise
Ichthyolaria densicostata = schwarze Dreiecke
Ichthyolaria squamosa = weiße Kreise

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at



©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Die drei ersten, unrotierten Faktoren brachten bei den einfach standardisierten Merkmalswerten eine Konfiguration der Individuen, wie sie in Abb. 20 dargestellt ist. Sie zeigt im vom ersten und zweiten Faktor aufgespannten Raum eine Überlappung der Arten *Ichthyolaria sulcata* (schwarze Kreise) und *Ichthyolaria densicostata* (schwarze Dreiecke), die jedoch durch Einbeziehung des dritten Faktors aufgehoben wird. Die drei Individuen der Art *Ichthyolaria squamosa* (weiße Kreise) sind innerhalb des Streuungsbereiches von *I. densicostata* gelegen und von dieser im dreidimensionalen Faktorenraum nicht zu trennen. Wenn noch keine Apriori-Information über die Arten vorhanden wäre, würde eine Klassifikation anhand der Konfiguration im Faktorenraum eine Zahl von sieben Fehlklassifikationen ergeben.

Ein sehr ähnliches Bild, aber doch im gewissen Maße differenziert, ergibt sich bei der Q-Mode-Faktorenanalyse, die an den komponentenspezifisch standardisierten Daten durchgeführt wurde (vgl. Abb. 21). Hier ist die Trennung zwischen *Ichthyolaria sulcata* und *Ichthyolaria densicostata* im Raum, der vom 1. und 2. Faktor aufgespannt wird, etwas deutlicher als bei den einfach standardisierten Merkmalswerten. Auch im dritten Faktor wird diese Trennung noch deutlicher als bei den einfach standardisierten Daten hervorgehoben. Besonders bemerkenswert ist jedoch, daß von den drei Individuen der Art *Ichthyolaria squamosa* zwei um vieles deutlicher von *Ichthyolaria densicostata* abgesetzt sind und als getrennte Gruppe aufscheinen. Es treten somit in dieser Konfiguration nur mehr vier Fehlklassifikationen auf.

Dies bedeutet, daß trotz annähernd optimaler Konfiguration der Individuen im Merkmalsraum, wie er in diesem Falle schon durch eine einfache Standardisierung erreicht wurde, die komponentenspezifische Standardisierung noch bessere, dem subjektiven Empfinden deutlich angenäherte Gruppierungen bringt.

Dank

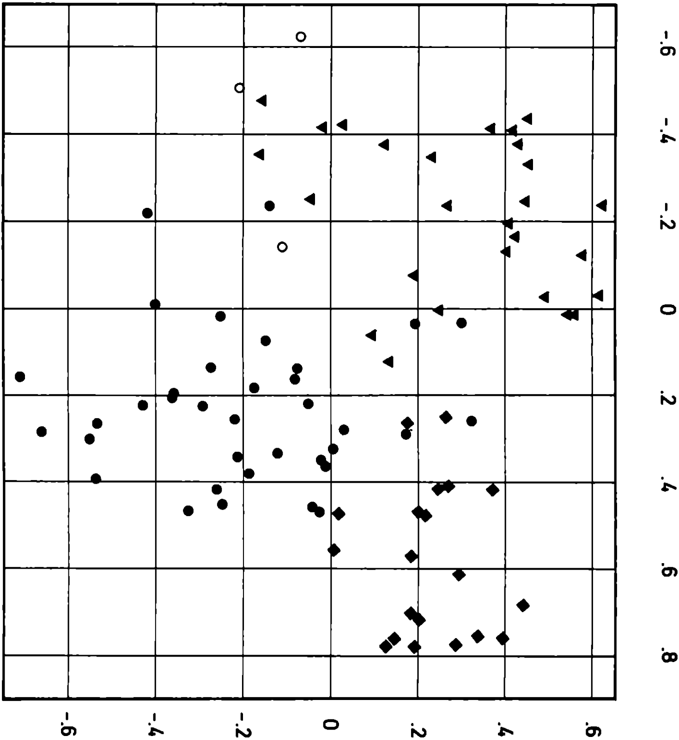
An erster Stelle möchte ich meinem Freund R. WYTEK (Rechenzentrum der Universität Wien) danken, der zahlreiche Denkanstöße gab und das Manuskript kritisch durchlas. Weiters danke ich Herrn Dr. R. BEIG (Institut für Theoretische Physik, Universität Wien) und Frau Th. HUBER für anregende Diskussionen bzw. für die Durchsicht des Manuskripts.

Die Arbeit wurde mit Mitteln des Theodor-Körner-Stiftungsfonds zur Förderung von Wissenschaft und Kunst durchgeführt.

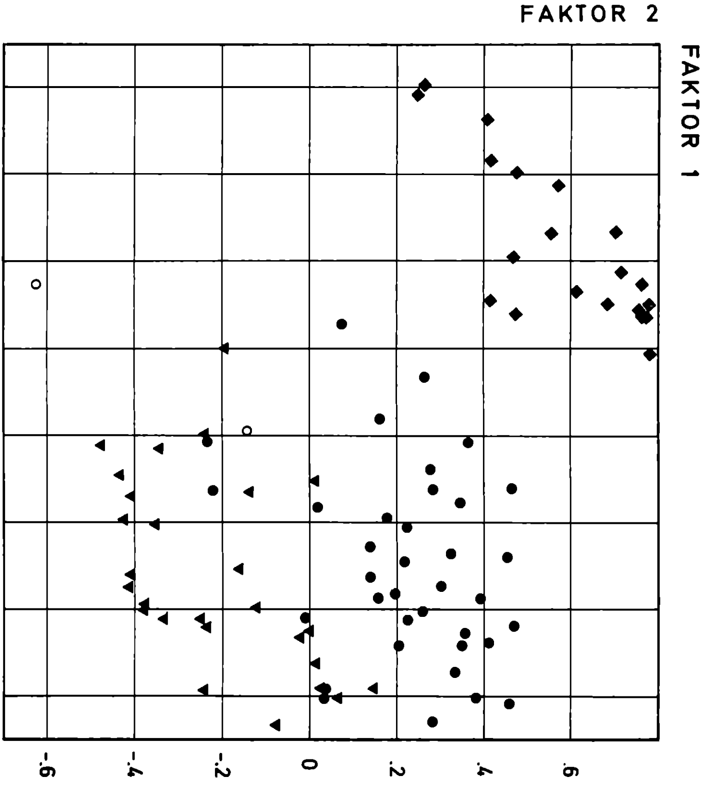
Abb. 21: Konfiguration von 86 Individuen der Gattung *Ichthyolaria* mit komponentenspezifisch standardisierten Merkmalswerten im dreidimensionalen Faktorenraum (Q-Mode-Analyse).
(Erläuterung der Signaturen siehe Abb. 20.)

©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

FAKTOR 3



©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at



©Akademie d. Wissenschaften Wien; download unter www.biologiezentrum.at

Literatur

AHRENS, H. J. (1974): Multidimensionale Skalierung. – 334 S., 40 Abb., Weinheim – Basel (Beltz).

BHATTACHARYA, C. G. (1967): A simple method of resolution of a distribution into Gaussian components. – Biometrics, *23*, 115–137, 4 Abb., Richmond (Virginia).

BLACKITH, R. E. & REYMENT, R. A. (1971): Multivariate Morphometrics. – 412 S., 44 Abb., London – New York (Academic Press).

BLACKWELDER, R. E. (1967): Taxonomy. – 698 S., 12 Abb., New York – London – Sydney (Wiley).

BOCK, H. H. (1974): Automatische Klassifikation. – 480 S., 42 Abb., Göttingen – Zürich (Vandenhoeck & Ruprecht).

BOX, G. E. P. & COX, D. R. (1964): An analysis of transformations. – J. Roy. Stat. Soc., Ser. B, *26*, 211–243, 8 Abb., London.

BOYCE, A. J. (1969): Mapping Diversity: A Comparative Study of Some Numerical Methods. – In: Numerical Taxonomy, 1–31, 10 Abb. Ed. A. J. Cole, London – New York (Academic Press).

BROLSMA, M. J. (1978): Benthonic Foraminifera. – Utrecht Micropaleont. Bull., *17*, 47–80, 25 Abb., Utrecht.

CATTELL, R. B. (1949): r_p and other coefficients of pattern similarity. – Psychometrika, *14*, 279–298, Richmond (Virginia).

CATTELL, R. B. (1965a): Factor analysis: An introduction to essentials. I. The purpose and underlying models. – Biometrics, *21*, 190–215, 2 Abb., Richmond (Virginia).

CATTELL, R. B. (1965b): Factor analysis: An introduction to essentials. II. The role of Factor analysis in research. – Biometrics, *21*, 405–435, 6 Abb., Richmond (Virginia).

CLIFFORD, H. T. & STEPHENSON, W (1975): An Introduction to Numerical Classification. – 229 S., 20 Abb., New York – San Francisco – London (Academic Press).

FANGMEYER, H. (1964): Die „Method of maximum likelihood" in der automatischen Klassifizierung. – Biometrische Z., *6*, 37–38, Berlin.

FERSCHL, F. (1978): Deskriptive Statistik. – 308 S., 43 Abb., Würzburg – Wien (Physica-Verlag).

GOODALL, D. W (1966): A new similarity index based on probability. – Biometrics, *22*, 882–907, Richmond (Virginia).

GOWER, J. C. (1966): Some distance properties of latent root and vector methods used in multivariate analysis. – Biometrika, *53*, 325–338, London.

GOWER, J. C. (1971): A general coefficient of similarity and some of its properties. – Biometrics, *27*, 857–871, Raleigh (North Carolina).

GREGOR, J. (1969): An algorithm for the decomposition of a distribution into Gaussian components. – Biometrics, *25*, 79–93, 2 Abb., Richmond (Virginia).

HARTIGAN, J. A. (1975): Clustering Algorithms. – 351 S., New York – London – Sydney – Toronto (Wiley).

HENNIG, W (1966): Phylogenetic Systematics. – 263 S., 69 Abb., Urbana – Chicago – London (University Illinois).

HOHENEGGER, J. (1980): Morphologische und taxonomische Analyse der liassischen berippten Ichthyolarien (Foraminifera). – Beitr. Paläont. Österreich, *7*, 17–117, 38 Abb., 6 Taf., Wien.

HOHENEGGER, J. (1981): *Ichthyolaria densicostata* n. sp., eine charakteristische Foraminifere des Unteren Lias Mitteleuropas. – Stuttgarter Beitr. Naturk., Ser. B, Nr. 74, 33 S., 8 Abb., 2 Taf., Stuttgart.

JARDINE, N. & SIBSON, R. (1971): Mathematical Taxonomy. – 286 S., 28 Abb., London – New York – Sydney – Toronto (Wiley).

JÖRESKOG, K. G., KLOVAN, J. E. & REYMENT, R. A. (1976): Geological Factor Analysis. – 178 S., 53 Abb., Amsterdam – Oxford – New York (Elsevier).

KREYSZIG, E. (1968): Statistische Methoden und ihre Anwendungen. – 422 S., 77 Abb., Göttingen (Vandenhoeck & Ruprecht).

KÜHN, W. (1976): Einführung in die multidimensionale Skalierung. – 186 S., 21 Abb., München – Basel (UTB-Taschenbücher).

MAYR, E. (1975): Grundlagen der zoologischen Systematik. – 370 S., 65 Abb., Hamburg – Berlin (Paul Parey).

MEDGYESSY, P. (1977): Decomposition of Superpositions of Density Functions and Discrete Distributions. – 308 S., 55 Abb., Bristol (Hilger).

ORLOCI, L. (1970): Automatic classification of plants based on information content. – Can. J. Bot., *48*, 793–802, 1 Abb., Ottawa.

RAUP, D. M. (1966): Geometric analysis of shell coiling: general problems. – J. Paleont., *40*, 1178–1190, 10 Abb., Tulsa (Oklahoma).

ROHLF, F. J., KISHPAUGH, J. & KIRK, D. (1977): NT-SYS. Numerical Taxonomy System of Multivariate Statistical Programs.

SCHNELL, P (1964): Eine Methode zur Auffindung von Gruppen. – Biometrische Z., *6*, 47–48, 1 Abb., Berlin.

SHEPARD, R. N. (1962a): The analysis of proximities: Multidimensional scaling with an unknown distance function. I. – Psychometrika, *27*, 125–140, 1. Abb., Richmond (Virginia).

SHEPARD, R. N. (1962b): The analysis of proximities: Multidimensional scaling with an unknown distance function. II. – Psychometrika, *27*, 219–246, 15 Abb., Richmond (Virginia).

SIMPSON, G. G. (1961): Principles of Animal Taxonomy. – XII + 247 S., 30 Abb., New York (Columbia University).

SNEATH, P. H. A. & SOKAL, R. R. (1973): Numerical Taxonomy. – 573 S., 81 Abb., San Francisco (Freeman).

SOLBRIG, O. T. & SOLBRIG, D. J. (1979): Introduction to Population Biology and Evolution. – 468 S., 175 Abb., Reading – Menlo Park – London – Amsterdam (Addison-Wesley).

SOKAL, R. R. & ROHLF, F. J. (1969): Biometry. – 776 S., 88 Abb., San Francisco (Freeman).

SOKAL, R. R. & SNEATH, P. H. A. (1963): Principles of Numerical Taxonomy. – 359 S., 39 Abb., San Francisco – London (Freeman).

TORGERSON, W S. (1958): Theory and methods of scaling. – 460 S., New York (Wiley).

VOGEL, F. (1975): Probleme und Verfahren der numerischen Klassifikation. – 410 S., 53 Abb., Göttingen (Vandenhoeck & Ruprecht).

White, M. J. D. (1978): Modes of Speciation. – 455 S., 31 Abb., San Francisco (Freeman).

Wishart, D. (1978): CLUSTAN. User Manual.

Zorn, H. (1972): Sind morphologische Merkmale von Organismen durch Verhältniszahlen quantitativ zu erfassen? – N. Jb. Geol. Paläont. Abh., *140*, 354–377, 17 Abb., Stuttgart.

Anschrift des Verfassers: Univ.-Doz. Dr. Johann Hohenegger, Institut für Paläontologie, Universität Wien, Universitätsstraße 7, A-1010 Wien.